CDA数据分析师系列丛书

医疗革命

医学数据挖掘的理论与实践

邵学杰 / 著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书以数据挖掘与模式识别的七大原理在临床医学中的运用案例为切入点，系统而全面地介绍了医学数据挖掘的基本方法与原理，对数据分析的常用算法进行了通俗易懂的讲解。本书最大的特色是采用了案例分析与实证的方法，每一个原理、算法都在案例讲解中生动地体现出来。更重要的是，本书对临床医学的数据挖掘与模式识别技术进行了开创性、系统性的讨论，用案例展现了数据挖掘技术如何与临床医学相结合，为广大的医生、医学数据挖掘工作者提供了很实用的技术示范、理念导入、系统思考。

本书所有概念的讲解基本结构为原理讲解与案例实操的二元结构，兼顾初学者与专业人士的需要。本书重点探讨了数据挖掘技术如何与临床医学深度融合，如何运用现代的数据挖掘理念、模式识别与机器学习的基本方法解决临床科研中的应用问题，为广大的科研型临床医生提供助力，为广大的数据分析人员找到行业应用的范例，为广大初学者提供努力学习的方向；更重要的是在这个大数据时代，我们可以亲自见证数据技术是如何改变并深刻影响着临床医学的科研与教学的。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目（CIP）数据

医疗革命：医学数据挖掘的理论与实践 / 邵学杰著. —北京：电子工业出版社，2016.9
（CDA 数据分析师系列丛书）
ISBN 978-7-121-29867-7

Ⅰ. ①医… Ⅱ. ①邵… Ⅲ. ①医学－数据采集－研究 Ⅳ. ①R-39

中国版本图书馆 CIP 数据核字(2016)第 211991 号

策划编辑：石　倩
责任编辑：石　倩
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×980　1/16　印张：12　字数：308 千字
版　　次：2016 年 9 月第 1 版
印　　次：2016 年 9 月第 1 次印刷
定　　价：49.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819　faq@phei.com.cn。

序

Big Data（大数据）在这几年突然火红于日常生活的各项领域中，连临床医疗也不例外，其实早期就存在许多通过数据来佐证或者分析预测结果的例子，但是当时在大多数的情况之下，统计运算不够快速成为了资料分析的一大限制，因此大多数资料是被临床研究工作者们放在一边而从未思考该如何运用的。伴随着信息科技的进步以及发达，能为我们所分析的数据将呈现爆炸性的成长，因此人们能从数据中学习的知识会更加丰富。和其他科学领域相比，需要透过大量临床试验的医学领域算是进步较缓慢的学科。管仲曾说过：“不明于数欲举大事，如舟之无楫而欲行于大海也。”意思是说在不清楚相关数据的情况下想做大事，无疑是没有桨的船想航行于汪洋大海中一样。也就是说，在医疗大数据的时代下，“dry lab”的医疗数据研究将会是协助医学领域快速进步的一大重要关键。本书通过大量临床医学的实例，由浅入深地介绍各项数据分析以及数据挖掘的方法和工具，将大量的临床医学数据化繁为简。相信无论是在校的学生或是临床研究者，本书都将会是学习或科研路上不可或缺的好伙伴。

谢邦昌
台北医学大学管理学院及大数据研究中心　院长/主任
中华市场研究协会理事长
中华资料采矿协会荣誉理事长

前　言

在医学大数据时代，数据技术带来了临床医学科研的革命性进步。本书通过对医疗数据挖掘的基本理论的阐述，将现代统计学与数据挖掘技术有机结合，讲述了大量的医学数据挖掘的案例，提供了大量的医学数据挖掘的实操方法。医学数据模式识别的七大原理与案例讲解是本书具有独创性的对医学数据技术的全面概括与总结，七大原理的首次提出也是医学数据挖掘技术上升到系统理论的重要实践与创新。无论是预测性建模、解释性建模、知识性建模与描述性建模，抑或是序列模式建模、依赖关系建模、异常模式建模，模式识别的类型规律跃然纸上，为专业人士或初学者厘清了数据挖掘与模式识别的基本类型特征。

不仅如此，本书选取的大量的医学数据挖掘案例为本书的实用性增加了学以致用的特色，凡认真阅读本书的读者都会从理论与实操两个层面全面、系统、实用地了解医学数据挖掘的原理与方法。本书以胰腺癌与二型糖尿病的关联规则、乳腺癌图片智能识别的挖掘算法、心电信号大数据的人工智能识别、低位前切保肛术的荟萃分析、贝叶斯网络预测高血压患者心血管风险、基线静息心率评估心血管事件、老年肺癌研究的荟萃分析等实用数据技术为切入点，使初学者能够掌握医学数据挖掘的基本理论与方法，因此是一本很好的入门级教科书。

对于资深的临床医生、医学博士、论文写作者而言，本书也是一本很好的案例参考书。特别是对于医学科研课题而言，本书提供了强大的实际操作技术培训与案例讲解，从顶级的国际期刊《自然》、《细胞》、《柳叶刀》等杂志选取经典的数据分析案例，用生动的方法让读者可以学到医学论文中数据、图表、算法的实际使用方法；因而对于专业人员而言，本书又是一本很好的资深级别的专业用书。

我们相信，无论您是初学者还是资深的专业人士，本书都将为您提供极大的可读性、趣味性和科学性。

目　录

第 1 章

数据分析与数据挖掘的力量

- 葡萄牙医生解决世界新生儿出生缺陷的故事
- 医学数据挖掘的主要定义
- 医学数据模式识别的七大原理与案例讲解
- 临床医学领域的机器学习与人工智能
- 多层神经元网络算法的基本原理

1.1 葡萄牙医生解决世界新生儿出生缺陷的故事

每年，全球大约有数以百万计的新生儿缺陷患者，原因包括遗传的、环境的、病毒性的，其中有高达 25%以上的新生儿先天缺陷找不到明确的原因。虽然超声医学、分子遗传检测技术已经有长足的进步，但依然有 8%左右的新生儿先天缺陷在世界某些地区找不到原因，葡萄牙医生用数据挖掘方法的解决方案对我们很有启发。

葡萄牙医生首先以全球各诊所新生儿出生记录数据为基础，包括出生年月日、性别、家庭住址三项基础统计数据，然后用空间地理信息做匹配关联分析，就是分析出生婴儿与空间地理位置的关联性，结果如图 1-1 和图 1-2 所示。

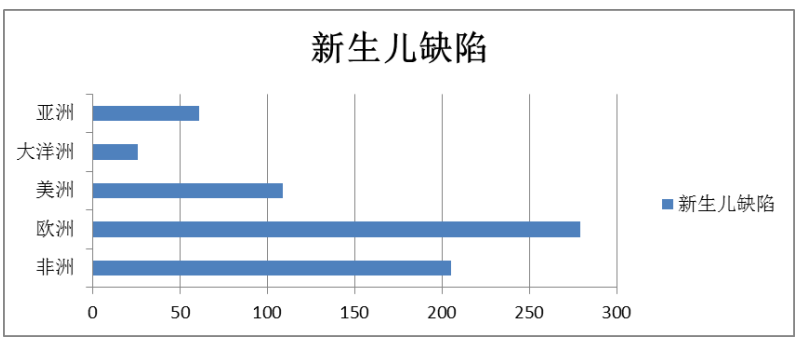


图 1-1　各大洲新生儿缺陷抽样分布数

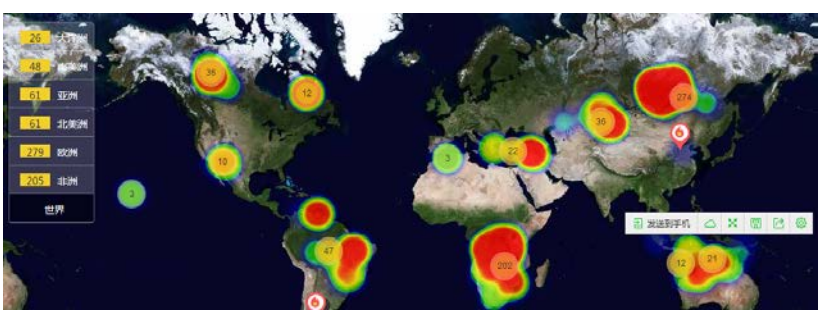


图 1-2　各大洲新生儿缺陷分布示意图

资料来源：葡萄牙医生 2014 年全球各大洲新生儿缺陷抽样调查报告

这项研究采用了最简单的单变量模型：变量是缺陷新生儿的出生地址，方法是采用全球抽样的均匀分布法，确保抽样数据的代表性。

抽样方法的正确性是指抽样的代表性和随机性，代表性反映样本与批质量的接近程度，而随机性反映检查批中单位产品被抽入样本纯属偶然，即由随机因素所决定。在对总体质量状况一无所知

的情况下，显然不能以主观的限制条件去提高抽样的代表性，抽样应当是完全随机的，这时采用简单随机抽样最为合理。在对总体质量构成有所了解的情况下，可以采用分层随机或系统随机抽样来提高抽样的代表性。在采用简单随机抽样有困难的情况下，可以采用代表性和随机性较差的分段随机抽样或整群随机抽样。这些抽样方法除简单随机抽样外，都是带有主观限制条件的随机抽样法。通常只要不是有意识地抽取质量好或坏的产品，尽量从批的各部分抽样，都可以近似地认为是随机抽样。

1. **单纯随机抽样**（simple random sampling）

将调查总体全部观察单位编号，再用抽签法或随机数字表随机抽取部分观察单位组成样本。

优点：操作简单，均数、率及相应的标准误计算简单。

缺点：总体较大时，难以一一编号。

2. **系统抽样**（systematic sampling）

该方法又称机械抽样、等距抽样，即先将总体的观察单位按某一顺序号分成 n 个部分，再从第一部分随机抽取第 k 号观察单位，依次用相等间距，从每一部分各抽取一个观察单位组成样本。

优点：易于理解、简便易行。

缺点：总体有周期或增减趋势时，易产生偏性。

3. **整群抽样**（cluster sampling）

总体分群，再随机抽取几个群组成样本，群内全部调查。

优点：便于组织、节省经费。

缺点：抽样误差大于单纯随机抽样。

4. **分层抽样**（stratified sampling）

先按对观察指标影响较大的某种特征，将总体分为若干类别；再从每一层内随机抽取一定数量的观察单位，合起来组成样本。有按比例分配和最优分配两种方案。

优点：样本代表性好，抽样误差减少。

以上四种基本抽样方法都属单阶段抽样，实际应用中常根据实际情况将整个抽样过程分为若干阶段来进行，称为多阶段抽样。

葡萄牙医生在本故事中采用了分群与分层抽样调查相结合的方法，按五大洲分群抽取，每个洲又按历史高发地区分层抽取。整群的聚类（cluster）是数据挖掘技术上一个很重要的概念，把某维度属性相近的实例聚类是数据技术最基础的方法；聚类后，距离太远的数据就是异常值。对数据处理的常规方法第一步就是聚类，把某些属性相近似的数据聚类后就可以进一步分析它们之间的关系，数据的聚类可以做回归（预测），数据的离散可以做预警（异常值）。

如图 1-3 所示，数据之间的关系可以从图形上表示出来，因此数据挖掘完全可以可视化地表现出来。就是说数据之间是有空间分布关系与距离的，用空间分布关系来表示数与数之间的关系，是现代数学的重要特征。

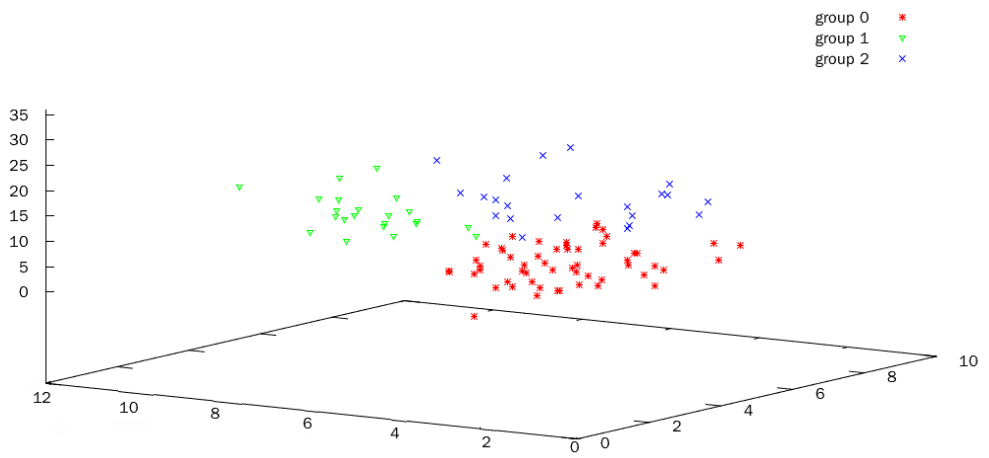


图 1-3　数据聚类图示效果

本故事中，葡萄牙医生的重要发现是：

① 欧洲大量新移民聚类中产生了新生儿缺陷高发的现象，这一数据甚至超过了传统落后地区非洲的新生儿出生缺陷率。

② 伊拉克战争、叙利亚战争、也门内战导致的难民大量涌入欧洲，人口的大规模迁徙改变了欧洲的新生儿人口健康状况。

就这样，葡萄牙医生用了一个简单的变量（婴儿出生地），代入了一个简单的分析框架——空间地理坐标与新生儿缺陷的关联性，用抽样方法获取数据，最后导出了近年来欧洲新生儿缺陷增加的主要原因：大规模移民难民潮。其中一个典型调查发现西班牙边境地区一个废弃的化学工厂是外来移民长期居住后新生儿缺陷发生的重要原因。

这是一个用数据进行知识发现（Knowledge-Discovery in Databases，KDD）的故事也是一个典型的流行病监测模型。数据库知识发现是数据挖掘最核心的意义。计算机时代，大量的数据被存放在数据库中，而不管是关系型数据库还是非关系型数据库，大量数据存储的成本都非常高昂；尤其在中国的三甲医院中，每天都有大量的门诊与住院数据产生，其中 80%是图像数据。一个普通的三甲医院每年产生大约 15TB～20TB 的新数据，这些数据中包含着许许多多疾病的新规律与知识发现，而用传统的统计学方法，用传统的手工或计算机方法已经无法处理或者无法准确地处理。这就是现代大数据技术产生的背景，包括传统的统计学、计算机技术、最优化分析技术、机器学习与人工智能、在线分析与检索技术。

1.2 医学数据挖掘的主要定义

1.2.1 数据挖掘的定义

数据挖掘(Data mining),又译为资料探勘、数据采矿。它是数据库知识发现(Knowledge-Discovery in Databases，KDD ）中的一个步骤。数据挖掘一般是指从大量的数据中通过算法搜索隐藏于其中信息的过程。数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法来实现上述目标。

数据挖掘利用了来自如下一些领域的思想：① 来自统计学的抽样、估计和假设检验；② 人工智能、模式识别和机器学习的搜索算法、建模技术和学习理论。数据挖掘也迅速地接纳了来自其他领域的思想，这些领域包括最优化、进化计算、信息论、信号处理、可视化和信息检索。一些其他领域也起到重要的支撑作用，尤其需要数据库系统提供有效的存储、索引和查询处理支持。源于高性能（并行）计算的技术在处理海量数据集方面常常是重要的。分布式技术也能帮助处理海量数据，并且当数据不能集中到一起处理时更是至关重要。

1.2.2 医学数据挖掘的故事

医学数据挖掘一般是指从大量的医学数据中通过算法搜索来认识隐藏于其中疾病新规律的过程。

今天这里要讲述一个关于肠道菌群与心血管疾病关联性的故事。在微生物学诞生后不久，人们就发现，在动物的消化道中存在有不少微生物。例如在牛、羊、兔等食草动物的胃或盲肠中，就存在大量以细菌为主的微生物群体。由于食草动物摄入的植食性饲料中，纤维素、半纤维素等多糖难以依靠动物体自身分泌的酶液消化，而微生物群体中包含的纤维素消化菌、半纤维素消化菌等可以较好地将多糖转化为低聚糖和寡糖，从而促进对这些营养物质的吸收。

随着医学的发展，人们也注意到，在人类的肠道，尤其是结肠（也就是平常所说的大肠）中，也存在着大量微生物。这些以细菌为主的微生物种类极多，数量极大。肠道菌群并非是生来就有的，它们实际上是“外来户”。在母体子宫内，胎儿所处的是一个几乎无菌的环境，因此胎儿肠道内也是无菌的。当胎儿出生之后的几天内，细菌通过分娩时阴道物质摄入、哺乳时的口腔摄入以及空气吸入等途径进入新生儿体内，并在肠道内定植，形成新生儿最初的肠道菌群。随着婴儿的成长，肠道菌群的种类结构逐渐趋于稳定，最终形成成熟的肠道菌群。这些微小的生物群体就这样不知不觉地定居到人体之内，悄无声息地与主人相随一生。

近期的多项研究表明，肠道菌群和人体的代谢疾病具有重要关系。肠道菌群失衡可能是造成肥胖、糖尿病等多种代谢异常的重要原因之一。造成代谢异常的主要原因，是失衡的肠道菌群产生的脂多糖等内毒素进入人体，被免疫细胞识别后产生多种炎症因子，使得机体进入低度炎症状态，从而产生代谢异常。例如，若长期进食高脂、高糖食物，可造成肠道菌群中条件致病菌比例增加，而

共生菌比例下降，从而使得食物中摄取的能量更容易转化为脂肪累积于皮下，造成肥胖。此外，低度炎症还能促使机体对胰岛素响应程度下降，造成胰岛素抵抗，进而发展为糖尿病。

这些医学观察的结论完全得益于数据挖掘技术的进步，医生们从医治经验中发现患有肠道疾病的人往往也同时患有心血管疾病。一开始医生们并没有注意到这个现象，当越来越多的病例记录了同一现象时，医生们开始怀疑两者之间的关联性。但是怀疑代替不了科学结论，需要定量化的数据支持，越来越多的病例数据汇总后经过关联规则算法最终找到了大量的支持病例，最终现代医学解开了这个秘密。肠道菌群与中风，原本风马牛不相及的两个病种终于确立了因果关系。

有意思的是，最新的医学数据挖掘表明，肠道菌群的数量分布居然与抑郁症有关联，医学科学家正在试图解开这个秘密。

这个故事生动地表达了医学数据挖掘的魅力与能量。利用大量的临床医学数据发现新的医学疾病规律正是数据挖掘在医学，特别是临床医学领域的巨大意义。

1.3 医学数据模式识别的七大原理与案例讲解

1.3.1 什么是模式识别

模式识别是指对表征事物或现象的各种形式的（数值的、文字的和逻辑关系的）信息进行处理和分析，以对事物或现象进行描述、辨认、分类和解释的过程，是信息科学和人工智能的重要组成部分。数据挖掘的本质就是模式识别。医学数据的七种模式识别方法分别是：

① 解释性数据建模；
② 描述性建模；
③ 预测性建模；
④ 知识性建模；
⑤ 序列模式建模；
⑥ 依赖关系的建模；
⑦ 异常与趋势建模。

建模就是建立模型，就是为了理解事物而对事物做出的一种抽象，是对事物的一种无歧义的书面描述。建立描述过程的性能的数学模型也称为建模，系统建模主要用于三个方面。①分析和设计实际系统。②预测或预报实际系统的某些状态的未来发展趋势，预测或预报基于事物发展过程的连贯性。③对系统实行最优控制。

数据挖掘中的建模，其本质就是模式识别的方法，包括数学定量描述与归因分析定性描述。医学模式识别就是利用临床医学大数据来建模，找到疾病之间的相互关系，无论是依赖关系，关联关系还是序列模式等关系都可以在数据中找到真相。

下面我们分别简述七个小故事来深入浅出地讲解这七个原理。

1.3.2　7 个小故事

1. 解释性数据建模

第一个小故事发生在朝鲜战争时期，第一、二次战役后以美国为首的联合国军遭遇了志愿军的重大打击，特别是长津湖一战，志愿军九兵团全线包围美骑一师，经过拼命抵抗美军好不容易才冲出重围。更有甚者，美军第八集团军司令沃克中将在给儿子的授勋途中车祸死亡，不得不由李奇微将军来接任。李奇微上任后的第一件事就是阅读美军的作战日志，他发现志愿军的每一次攻势都只有七天左右，他称之为“礼拜攻势”；经过深思熟虑，李奇微发现这是由于志愿军的补给只能维持七天所致。于是乎，美军依靠战场日志的“回放”发现了志愿军的弱点，美军制定了对付志愿军“礼拜攻势”的“磁性战术”，志愿军进攻时美军节节撤退避其锋芒，第七天开始发动反攻。这种战术给志愿军带来了很大的威胁，李奇微也成为志愿军很难对付的敌人。

这是一个典型的数据挖掘与模式识别案例。这里的数据记录就是美军的“战场日志”。通过对战场日志的梳理，李奇微发现了志愿军的攻击规律是七天一个周期，为了搞清楚原因，美军将领李奇微运用他的军事经验对“礼拜攻势”进行了很好的归因分析——志愿军是由于补给问题而导致的。这也是一个解释性建模的典型案例。

解释性建模是数据挖掘中的一个重要的模式识别方法。解释性建模的实质是模糊建模，模糊建模的概念由 Zadeh 提出后，在数据挖掘、模式识别、故障诊断、预测、监督与控制等方面得到了迅速的发展和应用，成为模糊理论与应用中重要的研究方向。模糊模型的特点在于它用模糊规则对知识进行表达，而且可以解决一些复杂的、非线性的、用传统的数学方法难以解决的问题。早期的模糊建模主要针对简单系统，采用总结专家经验的方式进行。

故事中的李奇微将军的战术就是典型的专家经验的合理运用。但是对于复杂系统，由于难以获得完备的专家经验，而数据相对容易获得，因此近年来基于数据的模糊建模成为研究的热点；但目前大多数研究将模糊模型作为一种函数逼近器，追求模糊模型对实际系统的拟合程度，即以模型的精确性为建模目标。

一个好的数学模型具备以下三点：① 描述性；② 预测性；③ 说明性。具体地说就是，一个好的数学模型能描述建模基于的系统，并且对其做出预测，同时能解释为什么这么建模以及建模得出的结论。针对以上三点，我们来看看数据和模型的区别。首先数据可以说是具有描述性，但仅是局部描述性，除非给出的数据能遍历每一种情况，而数学模型则具有全局描述性。其次，数据的预测性表现在可以通过数据建立模型，来给出预测结果。最后，好的数学模型能明确解释数据的走向，而光看数据你只能知道数据是怎么变化的，但不知道为什么这么变。建模和数据是相辅相成的，针对一个问题，建模是将其抽象到纯数学层面以寻求普适的解决方法与结论，数据是用来验证建模的结论，或者是辅助求解模型的（比如有些固定参数需要通过具体的实验或者观测数据来确定）。当然，只有用在好模型上，数据才会显得有意义。最后，如果数学建模真的因为大数据而没用了，也不会有那么多应用数学家还在探讨关于数学建模的问题。

2. 描述性建模

第二个小故事发生在微软公司的面试题中：如何在已知身高、体重、性别、年龄四个指标，但无法知道 18 个学生中任何一个人的照片及影像资料的条件下分析出这 18 个学生的身材数据?

描述性建模也是数学中常见的建模方法，其基本原理是从特殊到一般，即从分析事物的具体情况出发，经过数学语言的构建而得到一个可以具体描述事物特征的方法。描述性数学模型反映了从特殊到一般的认识过程，它是从分析客观事物的具体特征入手，经过逐步抽象而得到的。把客观事物中的关系概括于一个数学结构之中，是描述性数学模型的主要特征，也是解决问题的重要手段。

例如：乘法的交换律，如 $3\times4=4\times3$ 和 $15\times3.5=3.5\times15$，我们可以用 $a\times b=b\times a$ 来抽象表达，这就是一种描述性建模的思路。从特殊到一般的归纳与演绎，从具体特征到抽象表达，数学化的架构过程就是描述性建模的主要方法。

确立数与数之间的关系，可以采用代数方法或是几何方法，也可以采用代数加几何的解析几何方法。下面我们用坐标系的方法来直观地观察数与数的集合关系。数据如表 1-1 所示。

表 1-1　cop lot（height-weight-sex）不同性别下身高对于体重的散点

num	name	sex	age	height	weight
1	1 alice	f	13	56.5	84.0
2	2 becka	f	13	65.3	98.0
3	3 gail	f	14	64.3	90.0
4	4 karen	f	12	56.3	77.0
5	5 kathy	f	12	59.8	84.5
6	6 mary	f	15	66.5	112.0

attach(st)

如图 1-4 与图 1-5 所描述的那样，先以纵坐标为体重、横坐标为身高看看两者之间的关系。加入“性别”这个分组变量后，转换纵坐标为身高，横坐标为体重再次分析 18 个学生的数据。利用纵坐标、横坐标的变换来观察数据之间的关系是描述性建模的重要分析方法。描述性建模的主要方法就是从具体的特例中利用数据语言抽象概括出事物的特征：18 个同学中除去两个同学外（这两个同学一个瘦小，一个肥胖）其余同学身材匀称（身高与体重正相关）。

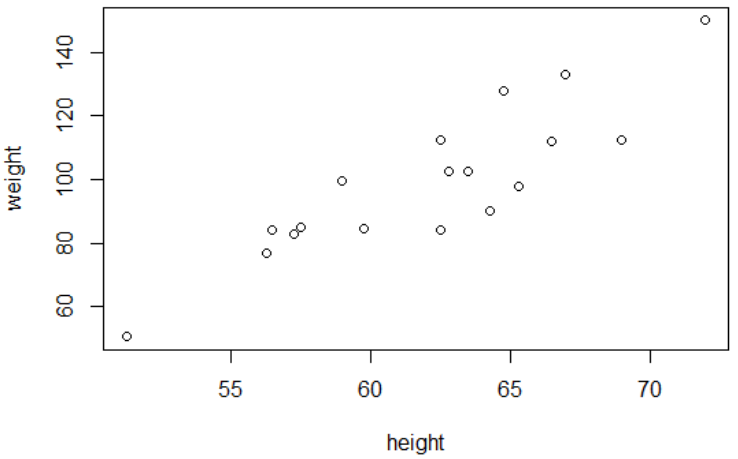


图 1-4　plot(height,weight)#身高对于体重的散点图

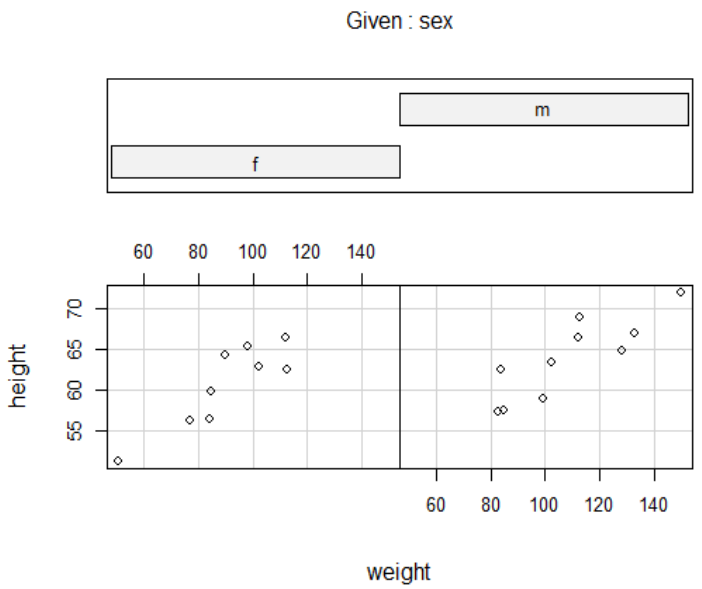


图 1-5　cop lot(height~weight|sex)#不同性别下身高对于体重的散点图

3. 预测性建模

第三个小故事是关于谷歌的大数据预测建模故事。如图 1-6 所示，谷歌公司依据全球用户对流感药物的在线查询情况已经可以提前六个月预测流感的爆发日期与流行路径。

每天都有成千上万的人通过 Google 来搜索信息，从旅途需要花费多长时间到怎样治疗他们孩子的病，各式各样的信息都有，这无疑极大地方便了人们的生活。

这一系列的搜索数据也从侧面显示出了搜索这些信息的人本身的情况，比如他们的想法、需求、忧虑等非常有价值的信息。如果这些信息的搜索可以准确地反映出人们的生存情况，那么分析人员就有可能利用这些信息追踪疾病情况，预测新商品的销售情况，甚至预测选举的结果。

图 1-6　谷歌全球流感趋势预测图

谷歌的研究人员探索了其中的可能性，并宣称他们能够根据人们在搜索引擎上留下的信息对流感进行“即时预报”。研究人员在《自然》杂志上撰文表示，能够进行这种预测的关键在于一旦人们患上了流感，就会在谷歌中搜索很多关于流感的相关信息，这就可以形成有关于流感流行情况的整体性趋势信号。

该文章还表示，如果把谷歌搜索引擎上的相关信息与美国疾病预防控制中心（Centers for Disease Control and Prevention，CDC）的流感监测信息进行调整对比，就可以提供更为精准的流感趋势预测；这不仅把人们在搜索引擎上留下的“垃圾”变成了拯救生命的“启示”，并比当前 CDC 的数据预测提早至少 2 周。

尽管谷歌的流感预测在后来三年的预测中准确度不高，引发了全球范围的争议，但是我们仍然要看到数据挖掘与机器学习在大数据领域的乐观前景。

小知识：回归就是预测吗？

19 世纪末，一场全新的革命已经蓄势待发。

英国统计学家道尔顿的一篇谈论人的身高的文章，提出了“回归”这个名词：“那些高个子的后代的身高，有种回归到大众身高的趋势。”道尔顿自己绝不会想到，自己竟然为这个世界创造了两个新的概念：一个是回归，一个是 regression towards the mean（趋中回归）。

19 世纪末的时候，古典概率理论已经比较成熟了，统计学诞生的基础已经有了。那时，一个叫卡尔·皮尔逊的年轻数学家，做了很多生物上、农业上的试验，使用了很多数据分析的方法，从最初的对数据的描述，到对数据的绘图，再到后来，使用拟合来寻找两组变量的联系……这个叫皮尔

逊的，就是现代统计学的鼻祖，也是第一次统计科学革命的领军人物。

皮尔逊 1904 来到伦敦大学学院（UCL），在他人生将尽的时候做了一件划时代的事情：建立了世界上第一个统计系。从此，统计学从数学中独立了，成为了另一个极具生命力的学科。

皮尔逊的儿子，继承了父业，也成了一代统计学大师。到这个时候，人们已经开始学会使用随机变量，使用概率模型来描述数据背后的那些不确定现象了。这一观念上的进步，使得回归问题有了新的发展。人们开始对回归问题进行新的解释，开始假设那些随机误差是怎样分布的；人们理所应当地选择了性质最好的那个分布函数：正态分布。如果假定这些误差都是期望为 0、方差一定、彼此不干扰的，那么，这就是“高斯同方差性回归模型”，即“经典回归模型”的雏形。人们又回到了最初的那些简单问题，使用直线来逼近数据，这就是 linear regression model（线性回归）的来历。

回归的意思就是有一个假设的或者说理论的线性或者非线性模型，然后通过回归的方法，将现有的数据向假设的模型拟合，无限地逼近。

对大数据的一般处理原则是“聚类则回归，离散则异常”。就是说如果一堆数据你不知道怎样处理，先做聚类分析，其中能够被聚类的数据就可以做回归（预测）分析，离散太大的数据可以看成离散值，就是异常值，通常用来预警。

4. 知识性建模

第四个小故事是大家耳熟能详的中国药物科学家屠呦呦获得诺贝尔奖的故事。这是一个典型的知识性建模，利用先验的知识经验，屠呦呦从中医古籍中找到了启发与灵感，先后筛选 2000 多种药物（在当时的条件下都是人工筛选），最后采用化学提纯与晶体分离的方法获得了青蒿素，为千百万疟疾患者带来福音。

大数据条件下的古方、古籍筛选可以采用高通量数据筛选的模型，用计算机技术，数据挖掘技术对千千万万的古方、古籍进行模式识别，利用先验的知识经验来探索发现新的知识规律。

5. 序列模式建模

第五个小故事就是著名的啤酒与尿布的故事。据说沃尔玛公司在应用数据挖掘技术分析销售数据的时候发现一个现象：啤酒销量增加，婴儿的尿布销量也相应地增加，这是为什么呢？经过分析他们发现年轻的爸爸们在为婴儿买完尿布后也会顺便给自己买一打啤酒，所以看似完全没有关联的两个商品品类建立起了联系。

这也是著名的“购物篮分析”，是数据挖掘常用的分析方法。

序列模式简单来说就是给定一个由不同序列组成的集合，其中，每个序列由不同的元素按顺序有序排列，每个元素（交易）由不同项目组成，同时给定一个用户指定的最小支持度阈值；序列模式挖掘就是找出所有的频繁子序列，即该子序列在序列集中的出现频率不低于用户指定的最小支持度阈值。

通过在多组商品中选取销量最高的几组进行关联分析，这时候往往能够发现一些意想不到的规律。商品排序的方法按时间就叫时间序列，按销量就叫销量序列，按品类就叫品类序列，可以假设 n 个序列，设定最小支持度阈值进行筛选。

6. 依赖关系建模

第六个小故事是医学上的关于激素的故事。长期的医学实践与数据分析表明，许多女性疾病与激素依赖高度相关，最常见的是子宫内膜癌与乳腺癌。重要的激素依赖性肿瘤是女性发病与死亡的重要原因。骨质疏松、冠心病也是女性多发的激素依赖性疾病。

2003 年，SAS 袭击中国，中国医生率先在世界上用激素抑制 SAS 病毒，取得了很好的疗效。然而，大量不规范的激素使用也使得幸存的 SAS 患者大多患有严重的骨质疏松，激素依赖疾病又一次被数据验证。

有趣的是，在医学界还没有解释清楚女性激素依赖疾病的发病原因和致病机理的时候，大量的激素依赖病例数据就已经确定了激素与这些女性疾病之间的关系，这些疾病的产生对激素水平有绝对的依赖，而不仅仅是共生关系。依赖与共生都是两个元素之间的关系，但有本质的不同，依赖是有 A 才有 B，共生是有 A 也有 B。

7. 异常与趋势建模

第七个小故事是关于中国医生应用离散值来预警并防范心血管疾病的故事。兰州大学医学院附属医院的医生在数据分析中应用离散值，特别是离散度的关系来判读心肌缺血事件的风险，具体的做法是利用心肌缺血患者心向量图 T 环改变与 Q_T 离散度的关系，观察 100 例心血管疾病患者心向量图 T 环异常指标及 Q_T 离散度。离散度越大，风险越高，超越阈值后就可以自动预警。

这里需要明确的是数据是有空间归属的，距离越远，离散度越大，风险越高。

1.4 临床医学领域的机器学习与人工智能

机器学习（Machine Learning，ML）是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科；专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。它是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域，它主要使用归纳、综合而不是演绎。

机器学习是一门让计算机在非精确编程下进行活动的科学。在过去十年，机器学习促成了无人驾驶车、高效语音识别、精确网络搜索及人类基因组认知的大力发展。机器学习如此无孔不入，你可能已经在不知情的情况下使用过无数次。这里我们向大家简述一个临床医学领域机器学习的案例。

乳腺癌的 X 射线诊断在临床医学中一直有着比较高的误诊率，为了解决这个问题，芝加哥大学的华裔医生尝试用机器学习的 SVM（支持向量机，一种机器学习的算法）来提高诊断的准确率。SVM 可应用于影像学诊断，故将自适应 SVM 应用于乳腺癌的诊断。从芝加哥大学放射科随机抽取了 200 例均得到病理证实的乳腺 X 片（104 例肿瘤 X 片中 46 例显示恶性，58 例显示良性），应用自适应 SVM 进行训练及检验，并把 SVM 的诊断结果同其他算法的诊断结果进行了对比，结果显示 SVM 诊断的准确性很高，如图 1-7 所示。

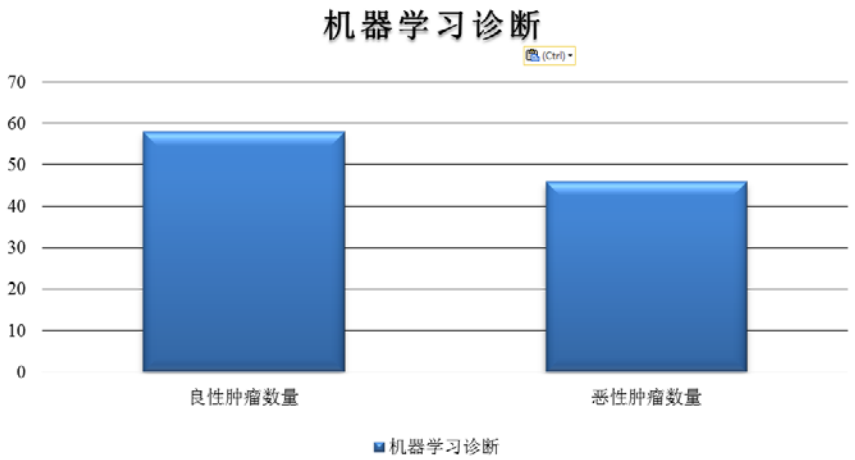


图 1-7　良性肿瘤数量与恶性肿瘤数量对比图

资料来源：《海南医学》2009 年第 20 卷第 9 期

自适应 SVM 的本质就是一种机器学习模仿人类大脑的举一反三机制，一般的运行原理是先给出一小部分样本数据学习，然后自动地推广到更大数据集合中，这些小样本就叫训练集。SVM 是一种有坚实理论基础的新颖小样本学习方法，它基本上不涉及概率测度及大数定律等，因此不同于以往的统计方法。从本质上看，它避开了从归纳到演绎的传统过程，实现了高效地从训练样本到测试样本的转导推理。SVM 算法明显优于人工读片产生的误诊率，表明了机器学习在乳腺癌的影像学初诊中产生的巨大作用，与人工判读相比，机器学习对乳腺癌诊断的准确率大大提高。

在临床工作中医生大多是凭借一些临床资料如患者的症状、体征以及各种检查结果根据临床经验得出结论，但是哪些资料的价值大，应该着重考虑，哪些只作次要考虑，各个医生的意见有时很不一致。究其原因还是个人的经验决定他们对各种资料所给的“权重”不同所致。人脑容量虽大但是对大样本量的资料的整合功能却较差，而 SVM 在这一方面有较强的优势，它能够通过小样本的学习最终获得诊断疾病的能力。

1.5 神经元网络的基本原理

神经元网络算法是基于模仿大脑神经网络的结构和功能而建立的一种信息处理系统。神经网络在一定学习规则下，对提供的学习样本进行学习，从中获取特征信息，并存储（记忆）在相应的权值及参数上。学习后，对于新的输入数据，网络可通过已获取的权值及参数，计算网络的输出。神经网络具有高度的非线性、容错性与自学习、自适应更新等功能，能够进行复杂的逻辑操作和非线性关系实现。

下面我们应用一个实例来说明神经元网络的运行原理。

案例：

为了对城市医疗能力进行评价，收集一批有代表性的城市医疗数据，评价指标为病床数、医生数、工作人员数、诊所数、死亡率，并给出了专家的评价结果，旨在建立评价城市的医疗建设绩效的模型，以便将其应用于评价任意城市的医疗建设绩效。收集数据见表 1-2 至表 1-4（单位：万人）。

其中：v——非常好，g——好，a——一般，b——差。

表 1-2　变数参数等级评价表

样　本	病床数	医生数	工作人员数	诊所数	死亡率	专家评价的医疗能力
上海	g	v	v	v	b	v
北京	a	v	v	v	g	v
沈阳	b	b	b	a	g	b
武汉	g	g	g	a	b	a
哈尔滨	v	g	a	b	a	a
重庆	g	g	b	b	b	b
成都	a	g	g	a	a	a

将取得的 7 个样本分别量化：定义 v、g、a、b 的取值为（1）v=1.5，g=0.5，a=–0.5，b=–1.5；也可以定义：v=3，g=1，a=–1，b=–3；v=6，g=2，a=–2，b=–6； v=10，g=7，a=4，b=1。由（1）定义可得上海等 10 个城市的样本取值，如表 1-3 所示。

表 1-3　各城市主要变量赋值表

样　本	病床数	医生数	工作人员数	诊所数	死亡率	专家评价的医疗能力	转换值	网络输出
上海	0.5	1.5	1.5	1.5	–1.5	1.5	0.9	0.8885
北京	–0.5	1.5	1.5	1.5	0.5	1.5	0.9	0.9581
沈阳	–1.5	–1.5	–1.5	–0.5	0.5	–1.5	0.1	0.1215
武汉	0.5	0.5	0.5	–0.5	–1.5	–0.5	0.37	0.38266
哈尔滨	1.5	0.5	–0.5	–1.5	–0.5	–0.5	0.37	0.369
重庆	0.5	0.5	–1.5	–1.5	–1.5	–1.5	0.1	0.1168
成都	–0.5	0.5	0.5	–0.5	–0.5	–0.5	0.37	0.34697
兰州	1.5	0.5	–0.5	0.5	1.5	1.5	0.9	0.8998
青岛	0.5	–1.5	1.5	1.5	–0.5	0.5	0.633	0.6419
鞍山	0.5	–0.5	–0.5	–1.5	1.5	0.5	0.633	0.6560

网络学习 35 万次后，网络收敛，总误差为 0.16，网络输出如表 1-4 所示，即存储网络学习后的有关权数与参数。

用学习后的网络，即可建立城市医疗能力评价模型。

表 1-4　各城市网络收敛值

样　本	病床数	医生数	工作人员数	诊所数	死亡率	网络输出	网络评价的医疗能力
天津	–1.5	0.5	–1.5	0.5	–0.5	0.122	b
广州	–0.5	0.5	0.5	0.5	–0.5	0.6687	g
南京	–1.5	0.5	0.5	0.5	–0.5	0.6423	g
西安	0.5	0.5	–0.5	0.5	0.5	0.6011	g
长春	0.5	0.5	0.5	–0.5	0.5	0.6333	g
太原	1.5	0.5	0.5	0.5	1.5	0.8851	v
大连	–1.5	–0.5	–1.5	–0.5	0.5	0.1134	b
济南	1.5	1.5	1.5	0.5	–0.5	0.8996	v
抚顺	0.5	–1.5	–1.5	–1.5	0.5	0.3869	a

这就是一个神经元网络的工作原理：先由人工赋值后进行训练集的运算，之后机器模仿训练集的方法进行自动化的运行。

学习是神经网络研究的一个重要内容，它的适应性是通过学习实现的。根据环境的变化，对权值进行调整，改善系统的行为。由 Hebb 提出的 Hebb 学习规则为神经网络的学习算法奠定了基础。Hebb 规则认为学习过程最终发生在神经元之间的突触部位，突触的联系强度随着突触前后神经元的活动而变化。在此基础上，人们提出了各种学习规则和算法，以适应不同网络模型的需要。有效的学习算法，使得神经网络能够通过连接权值的调整，构造客观世界的内在表示，形成具有特色的信息处理方法，而信息的存储和处则体现在网络的连接中。

根据学习环境不同，神经网络的学习方式可分为监督学习和非监督学习。在监督学习中，将训练样本的数据加到网络输入端，同时将相应的期望输出与网络输出相比较，得到误差信号，以此控制权值连接强度的调整，经多次训练后收敛到一个确定的权值。当样本情况发生变化时，经学习可以修改权值以适应新的环境。使用监督学习的神经网络模型有反传网络、感知器等。非监督学习时，事先不给定标准样本，直接将网络置于环境之中，学习阶段与工作阶段成为一体。此时，学习规律的变化服从连接权值的演变方程。非监督学习最简单的例子是 Hebb 学习规则。竞争学习规则是一个更复杂的非监督学习的例子，它是根据已建立的聚类进行权值调整。自组织映射、适应谐振理论网络等都是与竞争学习有关的典型模型。

如图 1-8 所示，神经元网络的实质是一个大型的分布式计算系统。人脑的神经元分布本来就是一个很好的分布式系统，每一个神经元做的信息处理工作只是一个信号的机械处理，但千千万万个神经元的机械动作叠加后就是智能。所有神经元的知识习得都是向外界学习、存储的结果。图 1-9 则描述了一个神经单元处理信息的三大步骤。

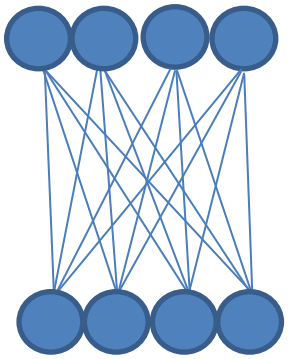


图 1-8　神经元交互示意图

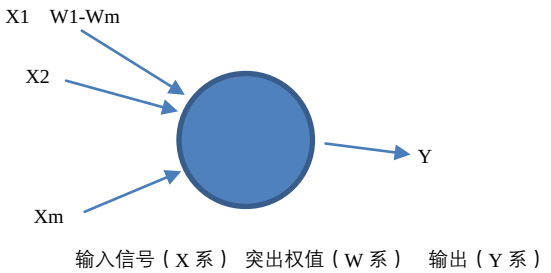


图 1-9　神经元信号处理流程图

本案例中的机器学习较好地模仿了人类大脑神经元的工作原理与方法。为了全面评估天津、广州、南京等 9 个城市的医疗水平，先用北京、上海、武汉等 10 个城市的数据作为训练数据，用床位、医护人员、诊疗机构数量、死亡率、专家评价六个变量作为输入信号，用 10 个城市的医疗经验作为突出权值（存量知识库），最后经过求和值的处理输出 Y 值。网络在学习后开始自动化处理 9 城市的赋值数据，最终收敛。这就是所谓的机器学习。

下面我们来看一个基于 Python 语言的神经元网络案例。

Frank Rosenblatt 第一个提出了感知器学习规则的概念。其主要思想是：定义一个算法去学习权重值 *w*，再将 *w* 乘以输入特征，以此来确定神经元是否受到了刺激。在模式分类中，可以应用这个算法确定样本属于哪一个类。将感知器算法置于机器学习的更广泛背景中：感知器属于监督学习算法类别，更具体地说是单层二值分类器。简而言之，分类器的任务就是基于一组输入变量，预测某个数据点在两个可能类别中的归属。感知器算法首先学习输入信号权值，然后得出线性决策边界，进而能够区分两个线性可分的类 1 和–1。

用 Python 实现感知器规则。在本节中，我们将用 Python 实现简单的感知器学习规则以分类鸢尾花数据集。请注意，为阐述清晰，这里省略了一些“安全检查”，若需要更“稳健”的版本，请参阅 Github 中的代码。

```
import numpy as np
class Perceptron(object):
def __init__(self, eta=0.01, epochs=50):
self.eta = eta
self.epochs = epochs
def train(self, X, y):
self.w_ = np.zeros(1 + X.shape[1])
self.errors_ = []
for _ in range(self.epochs):
errors = 0
for xi, target in zip(X, y):
```

```
    update = self.eta * (target - self.predict(xi))
    self.w_[1:] += update * xi
    self.w_[0] += update
    errors += int(update != 0.0)
    self.errors_.append(errors)
    return self
    def net_input(self, X):
    return np.dot(X, self.w_[1:]) + self.w_[0]
    def predict(self, X):
    return np.where(self.net_input(X) >= 0.0, 1, -1)
```

对于下面的示例，这里将从 UCI Machine Learning Repository 载入鸢尾花数据集，并只关注 Setosa 和 Versicolor 两种花。此外，为了可视化，这里将只使用两种特性：萼片长度（sepal length）和花片长度（petal length），如图 1-10 所示。

```
    import pandas as pd
    df = pd.read_csv('https://archive.ics.uci.edu/ml/machine-learning-
databases/iris/iris.data', header=None)
    # setosa and versicolor
    y = df.iloc[0:100, 4].values
    y = np.where(y == 'Iris-setosa', -1, 1)
    # sepal length and petal length
    X = df.iloc[0:100, [0,2]].values
    %matplotlib inline
    import matplotlib.pyplot as plt
    from mlxtend.evaluate import plot_decision_regions
    ppn = Perceptron(epochs=10, eta=0.1)
    ppn.train(X, y)
    print('Weights: %s' % ppn.w_)
    plot_decision_regions(X, y, clf=ppn)
    plt.title('Perceptron')
    plt.xlabel('sepal length [cm]')
    plt.ylabel('petal length [cm]')
    plt.show()
    plt.plot(range(1, len(ppn.errors_)+1), ppn.errors_, marker='o')
    plt.xlabel('Iterations')
    plt.ylabel('Missclassifications')
    plt.show()
    Weights: [-0.4 -0.68 1.82]
```

神经元网络的 X 系（信号系）、W 系（权值系）和 Y 系（输出系）凸显了数据挖掘技术的特点：计算机代码+数学函数。如图 1-10 所示，感知器的快速收敛是对算法是否有效最好的检验，两种花被有效地区分说明了函数建立的正确性与算法、代码运行的有效性。收敛最本质的意义是指算法函数有效地产生了结果，就像传统的线性方程有了根式解，就是函数存在可以逼近的极限，收敛是一切数据挖掘计算机数据处理追求的目标。

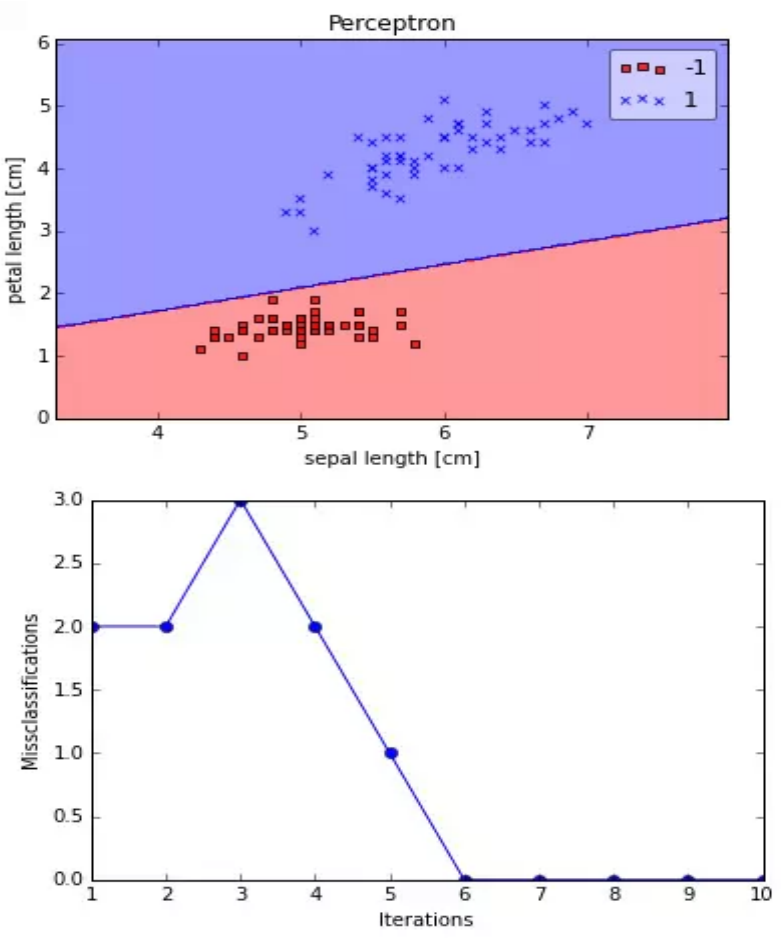


图 1-10 感知器收敛并完美区分出了这两种花

资料来源：英文出处（sebastianraschka.com）；转引自“Python 开发者”

小知识：统计学的诞生

统计学诞生于对国家的研究，特别是对其经济以及人口的描述。当时现代数学尚未形成。因此那时的统计史基本上属于经济史的范畴。

现代统计主要起源于研究总体（population）、变差（variation）和简化数据（reduction of data）。第一个经典文献属于 John Graunt（1620—1674），其具有技巧的分析指出了把一些庞杂、令人糊涂的数据化简为几个说明问题的表格的价值。他注意到在非瘟疫时期，一个大城市每年死亡数有统计规律，而且出生儿的性别比为 1.08，即每生 13 个女孩就有 14 个男孩。大城市的死亡率比农村地区要高。在考虑了已知原因的死亡及不知死亡年龄的情况下，Graunt 估计出了六岁之前儿童的死亡率，并相当合理地估计出了母亲的死亡率为 1.5%。因此，他从杂乱无章的材料中得出了重要的结论。他还给出了一个新的生命表。直到 1830 年，几乎所有的经验分布都是关于一维误差或一个非数值变量。在 1830 年之后，天文学家和社会学家 Adolphe Jacques Quetele（1796—1874）使得诸如身高体重之类度量值的变量的经验分布通俗化。他在生物统计研究中大量利用了理论二项分布和正态分布。后来 Ladislaus von Bortkiewicz（1868—1931）报告了在普鲁士兵团中由马踢造成的受伤事故，发现 Poisson（帕松）分布和官方统计学有关。在计算血红细胞数目上，Poisson 分布也被 Ernst Abbe（1840—1905）所用。从那时起，该分布被大量地用于计数的试验中，比如闪光的计数。

在生物学上，统计方法使得 Johann Gregor Mendel（孟德尔）（1822—1884）认识到某些主要遗传基因的存在，它们在 0、l 和 2 三个水平显现，其中水平 0（双隐性）能和水平 1 和 2 区别开来。他能确定有相同或不相同水平的个体之间交配的结果，而且提出了某些生物学事件等价于掷一个硬币的模型；他能对任意交配的结果给出概率并用实验来验证其假设。

Philippe Pinel（1745-1826）和 Pierre Charles alexandre Louis（1787—1872）开始了建立疾病分类的困难课题；这些工作人员保存了精确和完整的所有病例的记录，并且能给出和预后有关的统计数字。Louis 能够利用跟踪调查的方法反驳当时广泛滥用的放血疗法。

值得注意的是，现代的数据挖掘与传统的统计学并不是一回事，它们之间有重大的区别。数据挖掘技术是计算机科学、数理统计、大数据技术的边缘学科。虽然数据挖掘技术中大量地采用统计分析技术，知识发现与模式识别仍然是数据挖掘的主要目标。传统的统计学主要用于可以手工计算的数据，在大数据时代，数据挖掘面临的对象是手工计算很难实现或者效率低下的大数据，数据的存储、计算、清洗、处理都需要依赖计算机来完成，在这样的背景下，数据挖掘已经不同于统计学。

第 2 章

临床医学的数据挖掘

- 房颤与肾功能关联现象的故事
- 支持向量机的算法原理与应用
- 疾病规律与统计学革命
- 老年肺癌研究
- 临床医学与数据挖掘的边缘学科

2.1 房颤与肾功能关联现象的故事

心房颤动（简称房颤）是最常见的持续性心律失常。随着年龄增长房颤的发生率不断增加，75岁以上人群可达 10%。房颤时心房激动的频率达 300～600 次/分，心跳频率往往快而且不规则，有时候可达 100～160 次/分，不仅比正常人心跳快得多，而且绝对不整齐，心房失去有效的收缩功能。房颤患病率的增长还与冠心病、高血压病和心力衰竭等疾病的增长密切相关。

很多医生经过长期的医疗临床从大量的病例中感知到了肾功能也许和房颤有某些联系，虽然这时还停留在经验医学的层次，但这样的“专家经验”往往是数据挖掘与知识发现的开端；所以临床医生们只要运用数据挖掘技术来验证，这种专家的经验就有可能转化为新的医学知识发现。事实上，数据的统计与临床医学有先天的血肉联系。任何医生的个体经验和病人的个体经验必须有大样本的数据集合支撑才有归纳总结的必要与意义，在医学的归纳与演绎中，数据分析占据宏观的高地，个体的病例只是沧海一粟。一个不懂得数据技术的医生只能是一个手术工匠，反之，能够用数据来支持临床经验的医生可以乘上归纳与演绎的翅膀，飞向医学科学的高处。

房颤与肾功能的关联分析也是如此，如何把医生的感觉与经验上升为新的医学知识发现呢?

首先是对临床医学的一种灵感，房颤属于心率失常，在大多数的三甲医院属于心脏中心或心内科的范围，而肾功能属于肾内科。要找到两者的联系规律，必须要有数据作为支撑。

其次是建立模型与变量。反映肾功能的主要检查指标有以下几种：①内生肌酐清除率；②血尿素氮；③血肌酐；④血红蛋白和红细胞数；⑤尿比重；⑥尿渗透压；⑦尿酚红排泄试验等。其中以前 3 种最为重要。内生肌酐清除率、血尿素氮、血肌酐的指标主要反映肾小球的滤过功能，慢性肾衰的贫血是由肾实质损伤所导致的，其程度与肾功能的损害程度相平行。而尿比重、尿酚红排泄试验、尿渗透压是检查肾小管功能的主要指标，直接反映肾脏的浓缩功能。平常人们最关注的肾功能检查，泛指血尿素氮及肌酐浓度的测定值。

我们选择肾功能、肾小球滤过率、蛋白尿作为主要的变量来考查它们与房颤之间的敏感度。

案例：

慢性肾功能不全与心血管疾病的发病率密切相关，因此我们假定慢性肾功能不全也会增加房颤发生的风险，既往研究在该方面尚无明确结论。

方法与结果：我们对本中心 5 年来的 678 例慢性肾功能不全患者进行了回顾性分析，评估了肾功能、肾小球滤过率、蛋白尿与房颤发生率之间的关系。与个别的肾小球滤过率小于或等于 90ml/(min・1.73/m^2)的患者相比，当肾小球滤过率降低至 60～89、30～59 和 15～29 ml/(min・1.73/m^2)时，各自的房颤多变量危险比和 95%的可信区间分别是 1.3（1.1～1.6），1.6（1.3～2.1），3.2（2.0～5.0；P 值小于 0.0001）。

资料来源：李平/刘小燕. 慢性肾功能不全与房颤发病率的相关性研究. 第三军医大学全军心血管病研究所. 2012 年.

这个案例生动地反映了医学数据挖掘与知识发现的基本规律：首先观察细致，专业经验丰富；其次是建模准确，变量选择科学；最后是数据采集准确，清洗得当。

下面我们用一个完整的案例来说明数据挖掘技术是怎样发现房颤与肾功能障碍之间的关联规则的。

慢性肾功能不全（Chronic Renal Insufficiency，CRI）是临床上常见的病理生理状态，流行病学调查显示美国成人 CKD 的患病率约为 11%，我国北京市慢性肾功能不全的患病率为 18.7%。CRI 患者常存在心脏结构和功能的改变及内环境的异常，与房颤一样可以同时合并存在诸多危险因素，如高血压、糖尿病、缺血性心肌病、肥胖、代谢综合征等。国外研究报道，慢性肾脏病患者房颤的患病率高达 13%～23%，因此，研究慢性肾脏病患者房颤的发生与相关性具有重要流行病学和临床意义。

背景：心房颤动（以下简称房颤）是普通人群中最常见的持续性心律失常，其患病率在普通人群为 1%～8%。房颤是缺血性中风和死亡最大的独立预测风险因素之一。慢性肾功能不全（Chronic Renal Insufficiency，CRI）是最常见的临床病理状态，而且多存在心脏结构和功能异常，以及内部环境的失衡。房颤和慢性肾脏疾病（Chronic Kidney Disease，CKD）具有一些共同的危险因素（如高血压、糖尿病、心血管疾病、肥胖、代谢综合征）。虽然国外研究表明 CRI 患者房颤的患病率较高，但关于 CRI 患者房颤的患病率及其相关关系国内鲜有研究报道。

方法：754 例病人（63.3±16.4 年，53.1%为男性）被纳入研究，CRI 患者来自重庆医科大学附属第二医院、重庆市中山医院、重庆市第三人民医院 2006 年 1 月～2009 年 6 月期间的住院患者。心房颤动、房性心动过速（持续时间大于 1 分钟）和频率房性早搏（8000 次/24h）均由 12 导联心电图记录、24 小时动态心电图和既往确切病史判定，纳入统计分析时均视为房性心律失常。计算并组间比较年龄、性别以及 eGFR 各亚组房颤的患病率。多元 logistic 回归分析房颤和房性心律失常的横向相关性及相关因素。

结果：平均估计肾小球滤过率分别为 18.7±17.0 ml/（min·1.73 m^2），89.4%的研究对象 eGFR≤45 ml/（min·1.73 m^2），45.6%的为血液透析病人。大于 70 岁的患者占 25%，房颤和房性心律失常患病率分别为 18.8%、24.8%。与 eGFR≥45 ml/（min·1.73 m^2）患者相比，eGFR≤45 ml/（min·1.73 m^2）患者房颤和房性心律失常的患病率更高（20.4%和 16.0%，P =0.001）。多元 logistic 回归分析显示年龄、冠心病、糖尿病、心力衰竭病史和透析与房颤及房性心律失常显著相关。性别、吸烟、饮酒、体重指数、高血压、增大的左房和高脂血症与房颤及房性心律失常无显著相关性。eGFR≤45 ml/（min·1.73 m^2）可能也与房颤相关，但统计学意义并不显著（P = 0.117）。

结论：CRI 患者房颤的患病率较普通人群更高，年龄、冠心病、糖尿病、心力衰竭病史和透析与 CRI 患者房颤的发生显著相关。

资料来源：王骄. 慢性肾功能不全与房性心律失常的相关研究. 重庆医科大学附属第二医院. 2011 年硕士论文.

本案例采用的资料与方法如下所述。

（1）病例来源

收集重庆医科大学附属第二医院、重庆市中山医院、重庆市第三人民医院三所医院 2006 年 1 月~2009 年 6 月住院的慢性肾功能不全患者的住院病例。

慢性肾功能不全的诊断与分期纳入标准：年龄>18 岁，确诊为慢性肾功能不全患者。排除标准：甲亢性心脏病、瓣膜性心脏病、家族性房颤、肺源性心脏病。

（2）诊断与分期标准

慢性肾功能不全的诊断依据：定义为经过肾活检或检测损伤标记物证实的肾脏损伤或肾小球滤过率（glomerular filtration rate，GFR）< 60 ml/（min・1.73 m^2），持续时间≥3 个月。肾脏损伤的标志物包括蛋白尿、尿试纸条、尿沉渣异常或肾脏影像学检查异常。GFR 可通过肾脏病膳食改良试验（MDRD）公式和 Cockcroft- Gault（CG）公式推算。

CKD 分期：根据美国肾脏病基金会 K/DOQI 专家组对 CKD 的分期方法，见表 2-1 美国肾脏病基金会 K/DOQI 专家组 CKD 分期标准。

表 2-1　美国肾脏病基金会 K/DOQI 专家组 CKD 分期标准

CKD Staging standard of National Kidney Foundation K/DOQI

分　期	特　征	CFR 水平（ml/min）
1	已有肾损害，CFR 正常或稍增加	≥90
2	CFR 轻度降低	60～89
3	CFR 中度降低	30～59
4	CFR 重度降低	15～29
5	ESRD（肾衰竭）	<15（或透析）

备注：表 2-1 这类数据挖掘的要点是找出准确的数据来源。采用三所医院的 700 多例住院病人数据是第一步，如果是大数据量，还需要 ETL 去除数据的杂音。第二步建立变量模型，就是找到因变量与自变量的关系科目，这是医学数据挖掘的关键之处，它需要丰富的临床经验与医学知识才能有效地完成。第三步就是面对数据的聚类回归或离散预警。其中，在大数据条件下，大多采用人工智能与机器学习；在小数据条件下，更多地采用传统统计学的老方法，如参数设计、假设检验、P 值、T 值的置信度衡量等。研究对象的基本特征如表 2-2 所示。

表 2-2　研究对象的基本特征

basic clinical characters of study objects

年龄（y）	63.3（16.4）
性别（男）	400（53.1%）
吸烟史	59（7.8%）
饮酒史	53（7.0%）
高血压	609（80.8%）
糖尿病	199（26.4%）

续表

高脂血症	45（6.0%）
充血性心力衰竭	96（12.7%）
冠心病	214（28.4%）
eGFR（ml/[min1.73m^2]）	18.7（17.0）
eGFR≤45ml/（min1.73m^2）	679（89.4%）
CKD1	3（0.4%）
CKD2	25（3.3%）
CKD3	142（18.8%）
CKD4	130（17.2%）
CKD5	454（60.2%）
体重指数（kg/m^2）	26.78（2.30）
总胆固醇（mmol/L）	4.3 ± 1.4
血尿酸（mg/dL）	433.3 ± 144.0
左房前后径（cm）	36.6 ± 6.6
透析	344（45.6%）

均值 ± 标准差或 *n*（%） N=754

资料来源：王骄. 慢性肾功能不全与房性心律失常的相关研究. 重庆医科大学附属第二医院. 2011 年硕士论文.

如表 2-3 所示，70 岁和 60 岁以上的患者分别占 43.1%和 61.9%。所有入选患者中房颤的比例为 18.8%。eGFR≤45 ml/（min・1.73 m^2）的患者房颤的患病率为 20.0%，与 eGFR>45 ml/（min・1.73 m^2）患者相比，两者房颤的患病率具有显著性差异（20.0%和 8.0%，*P*=0.036）。随着年龄的增长房颤的患病率亦显著增加，在<40 岁、40～49 岁、50～59 岁、60～69 岁、70～79 岁、80 岁以上各组分别为 5.4%、11.1%、12.7.0%、19.7%、24.3%、28.6%（*P*<0.001）。男性和女性房颤的患病率无显著性差异（17.8%和 20.1%，*P*=0.419）。

表 2-3 性别、年龄、肾功能与心房颤动

	N=755	心房颤动 *n*（%）	*P* 值
总体	754	142（18.8%）	
eGFR（ml/[min1.73m^2]）			0.036
≤45	679	13.6（20.0%）	
≥45	75	6（8.0%）	
年龄（y）			p<0.001
<40	74（9.8%）	4（5.4%）	
40-49	63（8.4%）	7（11.1%）	
50-59	150（19.9%）	19（12.7.0%）	

续表

60-69	142（18.8%）	28（19.7%）	
70-79	206（27.3%）	50（24.3%）	
>80	119（15.8%）	34（28.6%）	
性别			0.419
男性	400	71（17.8%）	
女性	354	71（20.1%）	

如表 2-4 所示，所有入选患者中房性心律失常的比例为 24.8%。与 eGFR≥45ml/（min.1.73m^2）患者相比，eGFR≤45ml/（min.1.73m^2）的患者房性心律失常的患病率有明显升高的趋势（25.8%和 16.0%，*P*=0.063）。随着年龄的增长房性心律失常的患病率亦呈增高趋势，在<40 岁、40～49 岁、50～59 岁、60～69 岁、70～79 岁、80 岁以上各组分别为 6.8%、14.3%、17.3%、25.4%、29.1%、33.6%（*P*<0.001）。男性和女性房性心律失常的患病率无显著性差异（17.8%和 20.1%，*P*=0.419）。

表 2-4　性别、年龄、肾功能与房性心律失常

		房性心律失常 *n*（%）	*P* 值
总体	754	187（24.8%）	
eGFR（mL/[min1.73m^2]）			0. 063
≤45	679	175（25.8%）	
≥45	75	12（16.0%）	
年龄（y）			<0.001
<40	74	5（6.8%）	
40-49	63	9（14.3%）	
50-59	150	26（17.3%）	
60-69	142	36（25.4%）	
70-79	206	60（29.1%）	
≥80	119	40（33.6%）	
性别			0.026
男性	400	86（21.5%）	
女性	354	101（28.5%）	

将年龄、性别、体重指数（BMI）、吸烟史、饮酒史、高血压、高脂血症、冠心病史、心力衰竭、eGFR（≤45 ml/[min 1.73 m^2]）、透析、左房增大（≥36mm）以及糖尿病纳入多元 logistic 回归分析，探讨上述因素与罹患房颤风险的关系，如表 2-5、表 2-6 所示。年龄、冠心病、心力衰竭、糖尿病、透析与房颤和房性心律失常发生的相对风险增加密切相关。eGFR（≤45 ml/[min 1.73 m^2]）亦可能增加房颤发生的风险，但统计学意义并不显著（*P*=0.117）。性别、体重指数（BMI）、吸烟史、饮酒史、高血压、高脂血症以及左房增大并不增加房颤发生的风险。尽管上述单因素分析 eGFR≤45 ml/

（min 1.73 m^2）可能与罹患房颤有关（*P*=0.063），但多因素 Logistic 回归分析显示 eGFR≤45 ml/（min 1.73 m^2）（*P*=0.229）、性别、BMI、高脂血症因素、吸烟史、饮酒史、高血压、左房增大等并不增加罹患房性心律失常的风险。

表 2-5　多因素 Logistic 分析与房颤危险

	OR（95%CI）	*P* 值
年龄（y）	1.061（0.916-1.229）	0.004
性别（男）	1.005（.675-1.496）	0.982
吸烟史	0.521（.187-1.454）	0.213
饮酒史	0.907（.323-2.544）	0.853
体重指数（kg/m^2）	0.979（.899-1.067）	0.634
eGFr（ml/[min1.73m^2]）（egfr≤45ml/[min1.73m^2]）	2.088（.832-5.238）	0.117
高血压	0.890（.521-1.521）	0.670
高脂血症	1.739（1.111-2.618）	0.205
左房前后径	1.623（1.073-2.455）	0.274
充血性心力衰竭	3.581（2.222-5.771）	<0.001
糖尿病	1.706（1.111-2.618）	0.015
	OR（95%CI）	P 值
透析	1.623（0.739-2.507）	0.022
冠心病史	1.565（1.024-2.391）	0.039

资料来源：王骄. 慢性肾功能不全与房性心律失常的相关研究. 重庆医科大学附属第二医院. 2011 年硕士论文.

表 2-6　多因素 Logistic 分析与房性心律失常危险

	OR（95%CI）	*P* 值
年龄（y）	1.019（0.891-1.164）	0.005
性别（男）	1.289（0.886-1.877）	0.185
吸烟史	0.611（0.241-1.546）	0.298
饮酒史	0.914（0.358-2.377）	0.851
体重指数（kg/m^2）	0.956（0.884-1.035）	0.266
eGFR（ml/[min1.73m^2]）（egfr≤45ml/[min1.73m^2]）	0.921（0.721-1.191）	0.229
高血压	1.001（0.613-1.635）	0.996
高脂血症	1.617（0.753-3.469）	0.718
左房前后径	1.113（0.776-1.596）	0.562
充血性心力衰竭	3.671（2.301-5.858）	<0.001
糖尿病	1.682（1.132-2.499）	0.010

续表

透析	2.116（1.342-3.336）	0.001
冠心病史	1.882（1.274-2.778）	0.001

资料来源：王骄. 慢性肾功能不全与房性心律失常的相关研究. 重庆医科大学附属第二医院. 2011 年硕士论文.

表 2-7 比较了透析与非透析患者的基本临床特征。透析患者房颤和房性心律失常的患病率显著高于非透析患者（25.6%和 13.2%，32.8%和 9.5%，P<0.001）。透析患者 eGFR（ml/[min 1.73 m^2]）值显著低于非透析患者（8.7 ± 8.9 和 27.1 ± 17.6，P<0.001）。此外，透析患者并存高血压（87.2%和 77.3%，P<0.001）、糖尿病（31.4%和 21.5%，P<0.001）、冠心病史（29.9%和 26.3%，P<0.001）和心力衰竭（20.6%和 6.1%，P <0.001）的比例亦显著高于非透析患者。

表 2-7　透析与非透析患者的临床特征

	透析（n=344）	非透析（n=410）	P 值
年龄（y）	60.5 ± 16.5	65.6 ± 15.9	0.596
性别（男）	182（52.9%）	218（53.2%）	0.942
吸烟史	25（7.3%）	34（8.3%）	0.602
饮酒史	28（8.2%）	25（6.1%）	0.275
高血压	300（87.2%）	309（77.3%）	<0.001
糖尿病	111（31.4%）	88（21.5%）	<0.001
充血性心力衰竭	71（20.6%）	25（6.1%）	<0.001
冠心病	103（29.9%）	108（26.3%）	0.273
eGFR（ml/[min1.73m^2]）	8.7 ± 8.9	27.1 ± 17.6	<0.001
eGFR≤45ml/（min1.73m^2）	339（98.5%）	340（82.9%）	<0.001
体重指数（kg/m^2）	26.81 ± 2.29	26.75 ± 2.31	0.988
总胆固醇（mmol/L）	4.09 ± 1.24	4.45 ± 1.43	0.054
血尿酸（mg/dL）	424.3 ± 152.8	440.8 ± 135.9	0.369
心房颤动	88（25.6%）	54（13.2%）	<0.001
房性心律失常	113（32.8%）	72（9.5%）	<0.001

均值 ± 标准差或 n（%）

资料来源：王骄. 慢性肾功能不全与房性心律失常的相关研究. 重庆医科大学附属第二医院. 2011 年硕士论文.

（3）研究局限

本研究作为回顾性横断面研究，主要存在以下研究局限：首先，纳入研究的慢性肾功能不全患者仅有 754 例，且全部为住院患者，大部分患者 GFR≤45ml/（min 1.73 m^2），近一半的患者接受透析治疗，因此研究对象的肾功能差、病情重，多数为终末期肾脏病患者，对于将研究结果用于解释

慢性肾功能不全的早期患者需要谨慎。其次，本研究对于房性心律失常的认定仅仅来源于病例资料，不排除病例资料缺失、诊断不完整等因素存在，且多数患者未经 24 小时动态心电图确认，因此有可能低估房颤的患病率。尽管存在上述限制，本研究仍然对于慢性肾功能不全患者罹患房颤的风险做出了有益的研究，仍然具有一定的流行病学和临床意义，特别是对晚期肾脏疾病患者具有参考价值。

（4）结论

慢性肾功能不全患者房颤患病率显著高于普通人群，冠心病史、心力衰竭、糖尿病、透析与房颤和房性心律失常发生的相对风险增加密切相关，年龄、性别、体重指数（BMI）、吸烟史、高血压等并不显著增加房颤和房性心律失常风险。深入研究慢性肾功能不全患者与房颤的关系具有重要的临床和流行病学意义。

（5）研究方法小结

本案例采用了医学数据分析中最常见的 Logistic 多因素回归预测方法。Logistic 回归又称 Logistic 回归分析，是一种广义的线性回归分析模型，常用于数据挖掘、疾病自动诊断、经济预测等领域。例如，探讨引发疾病的危险因素，并根据危险因素预测疾病发生的概率等。以胃癌病情分析为例，选择两组人群，一组是胃癌组，一组是非胃癌组，两组人群必定具有不同的体征与生活方式等。因此因变量就为是否胃癌，值为“是”或“否”，自变量就可以包括很多了，如年龄、性别、饮食习惯、幽门螺杆菌感染等。自变量既可以是连续的，也可以是分类的。然后通过 logistic 回归分析，可以得到自变量的权重，从而可以大致了解到底哪些因素是胃癌的危险因素。同时根据该权值可以根据危险因素预测一个人患癌症的可能性。

函数 $y=f(x)$ 中，很难找到一个函数当自变量 x 发生变化时，y 值仅取两个值或有限值，这时，我们的解决方法是不直接分析 x 与 y 的关系，而是分析 y 取某个值的概率 P 与 x 之间的关系。用 P 等价替换 y，Logistic 回归有效解决了分类型变量值的问题，在医学数据分析中对疾病的危险因素分析有独特的作用，成为医学统计学中使用最多的回归分析方法。

OR 的含义与相对危险度相同，指暴露组的疾病危险性为非暴露组的多少倍。OR>1 说明疾病的危险度因暴露而增加，暴露与疾病之间为“正”关联；OR<1 说明疾病的危险度因暴露而减少，暴露与疾病之间为“负”关联。还应计算 OR 的置信区间，若区间跨 1，一般说明该因素无意义。

关联强度大致如表 2-8 所示。

表 2-8　OR 关联强度表

OR 值	联系强度
0.9-1.0　1.0-1.1	无
0.7-0.8　1.2-1.4	弱（前者为负关联，后者为正关联）
0.4-0.6　1.5-2.9	中等（同上）
0.1-0.3　3.0-9.0	强（同上）
<0.1　10.0 以上	很强（同上）

如表 2-6、表 2-7 显示，OR 值表明充血性心力衰竭与透析是最强的房颤关联因素，OR 值分别为 3.671 和 2.116。

Logistics 回归几乎是最有用的医学数据挖掘工具之一。它的三大用途包括：

① 寻找危险因素，正如上面所说的寻找某一疾病的危险因素等。

② 预测，如果已经建立了 Logistic 回归模型，则可以根据模型，预测在不同的自变量情况下，发生某病或某种情况的概率有多大。

③ 判别，实际上跟预测有些类似，也是根据 Logistic 模型，判断某人属于某病或属于某种情况的概率有多大，也就是看一下这个人有多大的可能性是属于某病。

回顾本案例的研究，我们发现房颤与肾功能障碍的关系发现是医学知识发现最典型的案例。首先是医生的临床经验“预感”到这两种疾病之间的关联性，其次是用统计学工具解决数据分析问题，用数据定量提出“不可辩驳”的事实，最后是归纳总结，产生新的医学知识发现。当代的中国医生面临的一个重要问题就是数据的搜集、清洗、分享很困难，每周末全国各地上演着数以百计的医学学术会议，绝大多数的医生在发布、分享他们的学术成果时普遍缺乏“不可辨驳”的数据分析，他们的结论仅仅是一些临床经验的分享，大部分缺乏有代表性的数据样本，更多的是引用国外的医学文献，数据病例获得、分享的困难极大地制约了中国临床医生科研水平的提高。

小知识：置信区间

置信水平是指总体参数值落在样本统计值某一区内的概率，一般用 1–α 表示；而置信区间是指在某一置信水平下，样本统计值与总体参数值间误差范围。置信区间越大，置信水平越高。

置信水平在抽样对总体参数作出估计时，由于样本的随机性，其结论总是不确定的。因此，采用一种概率的陈述方法，也就是数理统计中的区间估计法，即估计值与总体参数在一定允许的误差范围以内，其相应的概率有多大，这个相应的概率称作置信度。公路工程中保证率一般用 β 表示，显著性水平用 α 表示，α+β=1。

置信水平是描述 GIS 中线元素与面元素的位置不确定性的重要指标之一。置信水平表示区间估计的把握程度，置信区间的跨度是置信水平的正函数，即要求的把握程度越大，势必得到一个较宽的置信区间，这就相应降低了估计的准确程度。

95%置信区间（Confidence Interval，CI）：当给出某个估计值的 95%置信区间为［a，b］时，可以理解为有 95%的信心（Confidence）可以说样本的平均值介于 a 到 b 之间，而发生错误的概率为 5%。

有时也会说 90%，99%的置信区间，具体含义可参考 95%置信区间。

置信区间具体计算方式为：

（1）知道样本均值（M）和标准差（ST）时

置信区间下限：a=M–n×ST；置信区间上限：a=M+ n×ST。

当求取 90% 置信区间时 n=1.645；

当求取 95% 置信区间时 n=1.96；

当求取 99% 置信区间时 n=2.576。

（2）通过利用蒙特卡洛（Monte Carlo）方法获得估计值分布时

先对所有估计值样本进行排序，置信区间下限：a 为排序后第 lower%百分位值；置信区间上限：b 为排序后第 upper%百分位值。

当求取 90% 置信区间时 lower=5 upper=95；

当求取 95% 置信区间时 lower=2.5 upper=97.5；

当求取 99% 置信区间时 lower=0.5 upper=99.5。

当样本足够大时，（1）和（2）获取的结果基本相等。

2.2 支持向量机的算法原理与应用

2.2.1 一个故事的开场白

我们先在这里讲一个美国 51 岁士注册护的莫尼卡·朗尼的故事。两年多前，因被诊断为乳腺原位癌，她做了部分乳房的切除手术。不久前，她随男友从伊利诺伊州来到密歇根州，成为当地中西部地区医学中心由一名护士，并请这里的肿瘤专家丹尼斯西特林博士为她治疗乳腺癌。然而，据《纽约时报》报道，当她按常规预约到西特林的肿瘤学办公室时，得到的是一个难以置信的消息：她根本没有患癌症!

病理学家从她的一块活组织切片上诊断她患上了乳腺原位癌。之后，朗尼接受了外科手术，右侧乳房切除了一个高尔夫球大小的组织，还接受了 6 个星期的放射治疗。如今，她的新医生却肯定地告诉她：病理学家的诊断是错的，她从来就没有患上这种疾病，她所接受的手术、放射性治疗、药物……还有害怕，都是不必要的。从心理上讲，这太可怕了，她其实根本不需要走过那段经历。与绝大多数女性一样，朗尼认为乳房活组织切片是鉴别乳腺癌的黄金标准。然而，根据《纽约时报》对乳腺癌病例的调查，根据活组织切片检查来诊断是否为乳腺癌初期，其实相当困难，诊断的结果可能完全出错。有一本名为《最后诊断》的书就是讲病理科医生故事的，医生们在一个个的病例上都可以有不同的意见。过去 30 年中，乳房 X 线照相术和其他成像技术的进步表明，病理学家们必须借助于更小的乳腺组织来作出判断，有些甚至只有几粒盐般大小。根据医学报告和对医生、患者的采访，对病理学家来说，在良性肿瘤与早期乳腺癌之间作出判断是一个有相当挑战的领域。佛罗里达医学院病理学系主任 ShahlaMasood 表示过去 30 年中对原位癌的诊断一直充满争议、混乱，存在过度治疗与治疗不足等问题。《纽约时报》的文章指出，这类诊断错误问题说明“医学是不精确的科学”，依然是经验科学。

乳腺原位癌也被称为 0 阶段癌或非侵入性癌症，在 20 世纪 80 年代乳房 X 光检查广泛应用之前，这种癌症很难被检查出来。今天，美国每年有 5 万多位女性被诊断出是乳腺原位癌患者。不正常的细胞堆积在乳腺管道内，外科手术可以在其发展为侵人性癌症前将之去除。据估计，如果不接受治疗，那么一生中这些癌细胞有 30%的可能转化为侵入性癌症。考虑到乳腺病理学的准确性，美国病理学家协会表示，协会将启动一个服务于查验乳腺组织的病理学家的志愿认证项目，其中的要求之

一是病理学家必须每年检验 250 个乳腺病例。

乳腺癌的极高的误诊率激发了数据技术的快速发展，支持向量机（SVM）算法在乳腺癌的诊断中发挥了意想不到的作用。

2.2.2　支持向量机的主要特点

支持向量机堪称最有效的数据分类工具。主要的特点就是以极小的样本量也能够推演全体的数据算法。在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中。从此迅速地发展起来，已经在许多领域（生物信息学、文本和手写识别等）都取得了成功的应用。

支持向量机（SVM）中的一大亮点是在传统的最优化问题中提出了对偶理论，主要有最大最小对偶及拉格朗日对偶。

SVM 的关键在于核函数。低维空间向量集通常难于划分，解决的方法是将它们映射到高维空间。但这个办法带来的困难就是计算复杂度的增加，而核函数正好巧妙地解决了这个问题。也就是说，只要选用适当的核函数，就可以得到高维空间的分类函数。在 SVM 理论中，采用不同的核函数将导致不同的 SVM 算法。核函数将 *m* 维高维空间的内积运算转化为 *n* 维低维输入空间的核函数计算，从而巧妙地解决了在高维特征空间中计算的“维数灾难”等问题，从而为在高维特征空间解决复杂的分类或回归问题奠定了理论基础。维数灾难（Curse of Dimensionality）：通常是指在涉及向量的计算的问题中，随着维数的增加，计算量呈指数增长的一种现象。维数灾难在很多学科中都可以碰到，比如动态规划，模式识别等。

数据投射的升维（从一维到二维，从二维到三维，从 *n* 维到 *n*+1 维）是数学及统计理论的巨大进步，比如我们有一个一维的数据分布是如图 2-1 的样子，你想把它用一个直线来分开，你发现是不可能的，因为他们是间隔的。所以不论你画在哪，比如绿色竖线，都不可能把两个类分开。

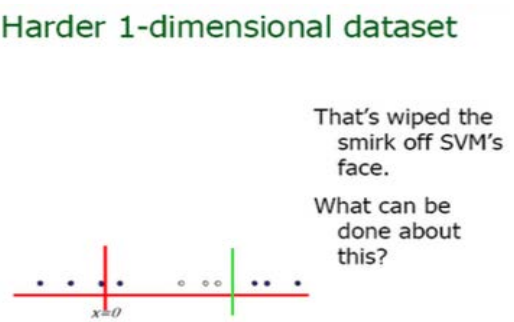


图 2-1　一维空间投射示意图

但是使用一个简单的升维的方法，把原来一维的空间投射到二维中，x→(x，x^2)。比如：

0→(0，0)

1→(1，1)

2→(2，4)

这时候就线性可分了，如图 2-2 所示。

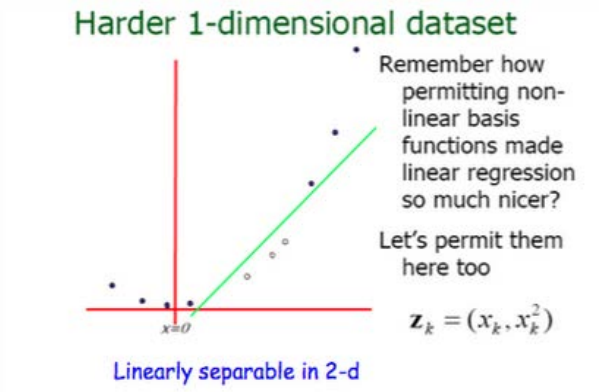


图 2-2　二维空间投射示意图

再举个例子，如图 2-3 所示，在一个二维平面里面，这样的情况是不可能只用一个平面来分类的，但是只要把它投射到三维的球体上，就可能很轻易地分类。

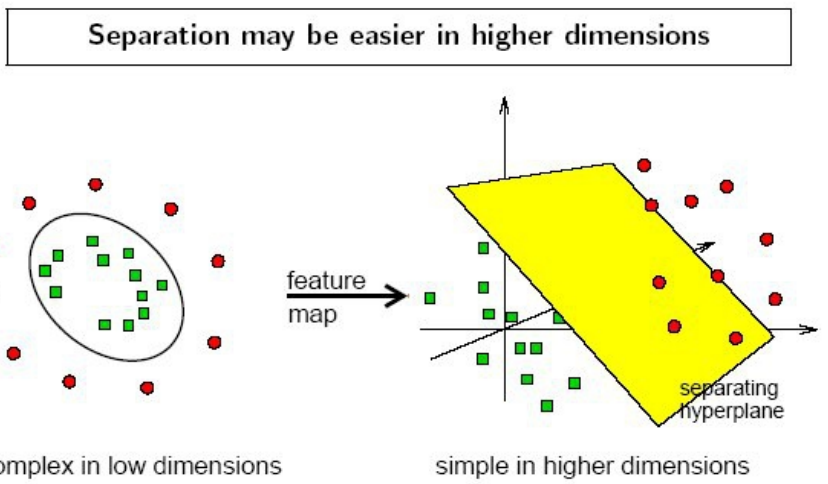


图 2-3　数的多维空间分界示意图

数学中，元素之间的关系很难分清的时候，我们往往把它投射到一个更高维的空间，在一个“超平面”上看到元素之间的分界线很清晰。比如，从欧式几何空间推广而来的希尔伯特空间就是如此。

在数学中，希尔伯特空间是欧几里得空间的一个推广，其不再局限于有限维的情形。与欧几里得空间相仿，希尔伯特空间也是一个内积空间，其上有距离和角的概念（及由此引申而来的正交性与垂直性的概念）。此外，希尔伯特空间还是一个完备的空间，其上所有的柯西序列等价于收敛序列，

从而微积分中的大部分概念都可以无障碍地推广到希尔伯特空间中。希尔伯特空间为基于任意正交系上的多项式表示的傅里叶级数和傅里叶变换提供了一种有效的表述方式，而这也是泛函分析的核心概念之一。希尔伯特空间是公式化数学和量子力学的关键性概念之一。一个抽象的希尔伯特空间中的元素往往被称为向量。在实际应用中，它可能代表了一列复数或是一个函数。例如在量子力学中，一个物理系统可以被一个复希尔伯特空间所表示，其中的向量是描述系统可能状态的波函数。详细的资料可以参考量子力学的数学描述相关的内容。量子力学中由平面波和束缚态所构成的希尔伯特空间，一般被称为装备希尔伯特空间（rigged Hilbert space）。

从数学的本质来看，最基本的集合有两类：线性空间（有线性结构的集合）、度量空间（有度量结构的集合）。

对线性空间而言，主要研究集合的描述，直观地说就是如何清楚地告诉别人这个集合是什么样子。为了描述清楚，就引入了基（相当于三维空间中的坐标系）的概念，所以对于一个线性空间来说，只要知道其基即可，集合中的元素只要知道其在给定基下的坐标即可。

对于很多分类问题，例如最简单的，一个平面上的两类不同的点，如何将它们用一条直线分开？在平面上我们可能无法实现，但是如果通过某种映射，将这些点映射到其他空间（比如说球面上等），我们有可能在另外一个空间中很容易找到这样一条所谓的“分隔线”，将这些点分开。

支持向量机（SVM）基本上就是这样的原理，但是 SVM 本身比较复杂，因为它不仅仅是应用于平面内点的分类问题。SVM 的一般做法是：将所有待分类的点映射到“高维空间”，然后在高维空间中找到一个能将这些点分开的“超平面”，这在理论上是被完全证明成立的，而且在实际计算中也是可行的。但是仅仅找到超平面是不够的，因为在通常的情况下，满足条件的“超平面”的个数不是唯一的。SVM 需要的是利用这些超平面，找到这两类点之间的“最大间隔”。为什么要找到最大间隔呢？我想这与 SVM 的“推广能力”有关，因为分类间隔越大，对于未知点的判断会越准确，也可以说是“最大分类间隔”决定了“期望风险”，总结起来就是：SVM 要求分类间隔最大，实际上是对推广能力的控制。支持向量机可用来做分类和拟合。其中分类的基本原理就是不仅仅要将分类点正确区分，而且还要使得分隔的距离最大，这便可以转化为凸二次规划问题来求解。

支持向量机（SVM）运用了泛函的方法，采用小样本也可以代表大数据的路径，用非线性的方法解决元素集合之间的关系问题，在医学上也有广泛的运用。

小结一下可以得出：支持向量机（Support Vector Machies）是由 Vapnik 等人于 1995 年提出来的。之后随着统计理论的发展，支持向量机也逐渐受到了各领域研究者的关注，在很短的时间就得到很广泛的应用。支持向量机是建立在统计学习理论的 VC 维理论和结构风险最小化原理基础上的，利用有限的样本所提供的信息对模型的复杂性和学习能力两者进行了寻求最佳的折衷，以获得最好的泛化能力。SVM 的基本思想是把训练数据非线性地映射到一个更高维的特征空间（Hilbert 空间）中，在这个高维的特征空间中寻找到一个超平面使得正例和反例两者间的隔离边缘被最大化。SVM 的出现有效地解决了传统的神经网络结果选择问题、局部极小值、过拟合等问题。并且在小样本、非线性、数据高维等机器学习问题中表现出很多令人嘱目的性质，被广泛地应用在模式识别之中。

支持向量机（SVM)也是一种二类分类模型。它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机。支持向量机还包括核技巧，这使它成为实质上的非线性分

类器。支持向量机的学习策略就是间隔最大化，可形式化为一个求解凸二次规划（convex quadratic programming）的问题，也等价于正则化的合页损失函数（后面也有解释）的最小化问题。支持向量机的学习算法是求解凸二次规划的最优化算法。支持向量机学习方法包含构建由简至繁的模型：线性可分支持向量机（linear support vector machine in linearly separable case）、线性支持向量机（linear support vector machine）及非线性支持向量机（non-linear support vector machine）。简单模型是复杂模型的基础，也是复杂模型的特殊情况。当训练数据线性可分时，通过硬间隔最大化（hard margin maximization），学习一个线性的分类器，即线性可分支持向量机，又称为硬间隔支持向量机；当训练数据近似线性可分时，通过软间隔最大化（soft margin maximization），也学习一个线性的分类器，即线性支持向量机，又称为软间隔支持向量机；当训练数据线性不可分时，通过使用核技巧（kernel trick）及软间隔最大化，学习非线性支持向量机。

小知识：支持向量机（SVM）的数学推导

1. 基本推导

Logistic 回归目的是从特征学习出一个 0/1 分类模型，而这个模型是将特性的线性组合作为自变量，由于自变量的取值范围是负无穷到正无穷，因此，使用 logistic 函数（或称作 sigmoid 函数）将自变量映射到（0,1）上，映射后的值被认为是属于 $y=1$ 的概率。

形式化表示就是假设函数

$$h_\theta(x)=g(\theta^T x)=\frac{1}{1-e^{-\theta^T x}}$$

其中 x 是 n 维特征向量，函数 g 就是 logistic 函数。

$g(z)=\frac{1}{1+e^{-z}}$ 的图像如图 2-4 所示。

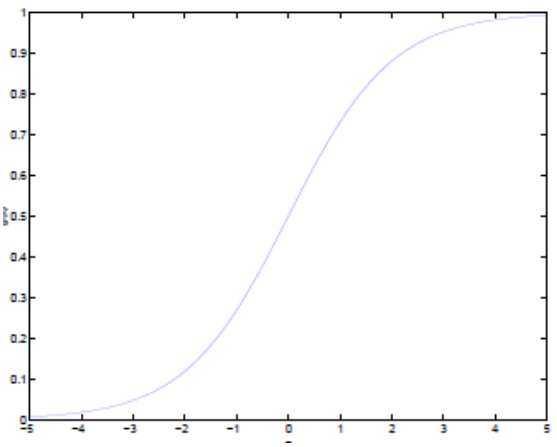


图 2-4　logistic 函数图像

可以看到，将无穷映射到了（0，1）。而假设函数就是特征属于 y=1 的概率。

$$P(y=1|x;\theta)=h_\theta(x)$$
$$P(y=1|x;\theta)=1-h_\theta(x)$$

当要判别一个新来的特征属于哪个类时，只需求 $h_\theta(x)$，若大于 0.5 就是 y=1 的类，反之属于 y=0 类。再审视一下 $h_\theta(x)$，发现 $h_\theta(x)$ 只和 $\theta^T x$ 有关，$\theta^T x>0$，那么 $h_\theta(x)<0.5$，$g(z)$只不过是用来映射，真实的类别决定权还在 $\theta^T x$。还有当 $\theta^T x \gg 0$ 时，$h_\theta(x)=1$，反之 $h_\theta(x)=0$。如果我们只从 $\theta^T x$ 出发，希望模型达到的目标无非就是让训练数据中 y=1 的特征 $\theta^T x \gg 0$，而 y=0 的特征 $\theta^T x \ll 0$。Logistic 回归就是要学习得到 θ，使得正例的特征远大于 0，负例的特征远小于 0，强调在全部训练实例上达到这个目标。

2. 图形化表示

图形化表示如图 2-5 所示。

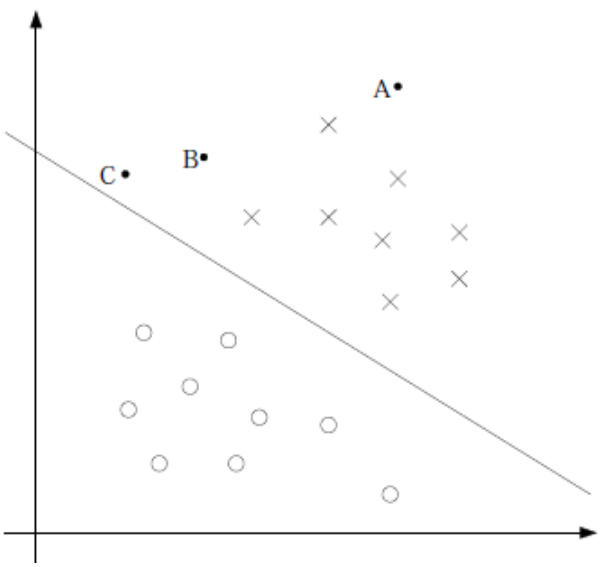


图 2-5　logistic 的数族分界示意图

中间那条线是 $\theta^T x$ =0，logistic 回归强调所有点尽可能地远离中间那条线。考虑上面 3 个点 A、B 和 C。从图中我们可以确定 A 是×类别的，然而 C 我们是不太确定的，B 还算能够确定。这样我们可以得出结论，我们更应该关心靠近中间分割线的点，让它们尽可能地远离中间线，而不是在所有点上达到最优。因为那样的话，要使得一部分点靠近中间线来换取另外一部分点更加远离中间线。我想这就是支持向量机的思路和 logistic 回归的不同点，一个考虑局部（不关心已经确定远离的点），一个考虑全局（已经远离的点可能通过调整中间线使其能够更加远离）。这是我的个人直观理解。

3. 形式化表示

我们这次使用的结果标签是 $y=-1$，$y=1$，替换在 logistic 回归中使用的 $y=0$ 和 $y=1$。同时将 θ 替换成 w 和 b。以前的 $\theta^T x=\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_n x_n$，其中认为 $x_0=1$。现在我们替换 θ_0 为 b，后面替换 $\theta_1 x_1+\theta_2 x_2+\cdots+\theta_n x_n$ 为 $w_1 x_1+w_2 x_2+\cdots+w_n x_n$（即 $w^T x$）。这样，我们让 $\theta^T x=w^T x+b$，进一步 $h_\theta(x)=g(\theta^T x)=g(w^T x+b)$。也就是说除了 y 由 $y=0$ 变为 $y=-1$，只是标记不同外，与 logistic 回归的形式化表示没区别。再明确下假设函数

$$h_{w,b}(x)=g(w^T x+b)$$

上一节提到过我们只需考虑 $\theta^T x$ 的正负问题，而不用关心 $g(z)$，因此我们这里将 $g(z)$做一个简化，将其简单映射到 $y=-1$ 和 $y=1$ 上。映射关系如下：

$$g(z)=\begin{cases}1, & z\geqslant 0\\ -1, & z<0\end{cases}$$

4. 函数间隔（functional margin）和几何间隔（geometric margin）

给定一个训练样本$(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})$，x 是特征，y 是结果标签。i 表示第 i 个样本。我们定义函数间隔如下：

$$r^{(i)}=y^{(i)}(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}+b)$$

可想而知，当 $y^{(i)}=1$ 时，在我们的 g(z)定义中，$w^T x^{(i)}+b\geqslant 0$, $\hat{y}^{(i)}$ 的值实际上就是 $\left|w^T x^{(i)}+b\right|$。反之亦然。为了使函数间隔最大（更大的信心确定该例是正例还是反例），当 $y^{(i)}=1$ 时，$w^T x^{(i)}+b$ 应该是个大正数，反之是个大负数。因此函数间隔代表了我们认为特征是正例还是反例的确信度。

继续考虑 w 和 b，如果同时加大 w 和 b，比如在$(w^T x^{(i)}+b)$前面乘个系数 2，那么所有点的函数间隔都会增大二倍，这个对求解问题来说不应该有影响，因为我们要求解的是 $w^T x^{(i)}+b=0$，同时扩大 w 和 b 对结果是无影响的。这样，我们为了限制 w 和 b，可能需要加入归一化条件，毕竟求解的目标是确定唯一一个 w 和 b，而不是多组线性相关的向量。这个归一化一会再考虑。

刚刚我们定义的函数间隔是针对某一个样本的，现在我们定义全局样本上的函数间隔

$$\hat{\gamma}=\min_{i=1,\cdots,m}\hat{\gamma}^{(i)}$$

其实就是在训练样本上分类正例和负例确信度最小那个函数间隔。

接下来定义几何间隔，先看图 2-6。

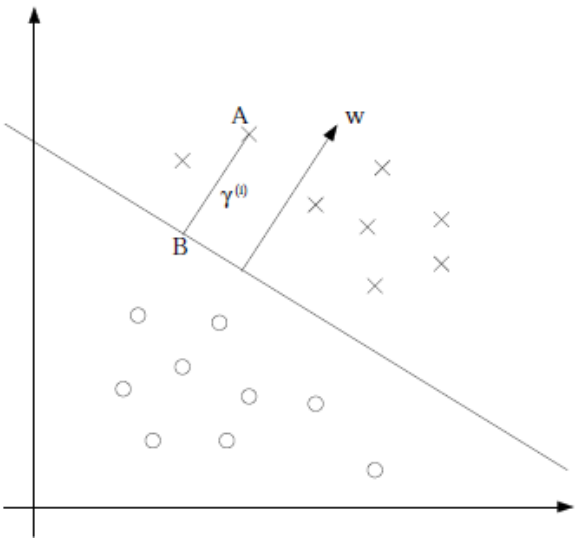


图 2-6　向量的投影梯度示意图

假设我们有了 B 点所在的 $w^T x+b=0$ 分割面。任何其他一点，比如 A 到该面的距离以 $\gamma^{(i)}$ 表示，假设 B 就是 A 在分割面上的投影。我们知道向量 BA 的方向是 w（分割面的梯度），单位向量是 $\frac{w}{\|w\|}$。A 点是 $(x^{(i)}, y^{(i)})$，所以 B 点是 $x=x^{(i)}-y^{(i)}\frac{w}{\|w\|}$（利用初中的几何知识），带入 $w^T x+b=0$ 得：

$$w^T\left(x^{(i)}-\gamma^{(i)}\frac{w}{\|w\|}\right)+b=0$$

进一步得到

$$\gamma^{(i)}=\frac{w^T x^{(i)}+b}{\|w\|}=\left(\frac{w}{\|w\|}\right)^T x^{(i)}+\frac{b}{\|w\|}$$

$\gamma^{(i)}$ 实际上就是点到平面的距离。

再换种更加优雅的写法：

$$\gamma^{(i)}=y^{(i)}\left(\left(\frac{w}{\|w\|}\right)^T x^{(i)}+\frac{b}{\|w\|}\right)$$

当 $\|w\|=1$ 时，不就是函数间隔吗？是的，前面提到的函数间隔归一化结果就是几何间隔。他们为什么会一样呢？因为函数间隔是我们定义的，在定义的时候就有几何间隔的色彩。同样，同时扩大 w 和 b，w 扩大几倍，$\|w\|$ 就扩大几倍，结果无影响。同样定义全局的几何间隔 $\hat{\gamma}=\min_{i=1,\cdots,m}\hat{\gamma}^{(i)}$。

5 最优间隔分类器（optimal margin classifier）

回想前面我们提到的目标是寻找一个超平面，使得离超平面比较近的点能有更大的间距。也就是我们不考虑所有的点都必须远离超平面，而关心求得的超平面能够让所有点中离它最近的点具有最大间距。形象地说，我们将上面的图看作是一张纸，我们要找一条折线，按照这条折线折叠后，离折线最近的点的间距比其他折线都要大。形式化表示为：

$$\begin{aligned}&\max_{\gamma,w,b}\ \gamma\\&\text{s.t. } y^{(i)}(w^Tx^{(i)}+b)\geqslant\gamma,\ i=1,\cdots,m\\&\|w\|=1\end{aligned}$$

这里用 $\|w\|=1$ 规约 w，使得 w^Tx+b 是几何间隔。

到此，我们已经将模型定义出来了。如果求得了 w 和 b，那么来一个特征 x，我们就能够分类了，称为最优间隔分类器。接下来的问题就是如何求解 w 和 b 的问题了。

由于 $\|w\|=1$ 不是凸函数，我们想先处理转化一下，考虑几何间隔和函数间隔的关系，$Y=\frac{\hat{\gamma}}{\|w\|}$，我们改写一下上面的式子：

$$\begin{aligned}&\max_{\gamma,w,b}\ \frac{\hat{\gamma}}{\|w\|}\\&\text{s.t.}\quad y^{(i)}(w^Tx^{(i)}+b)\geqslant\gamma,\ i=1,\cdots,m\end{aligned}$$

这时候其实我们求的最大值仍然是几何间隔，只不过此时的 w 不受 $\|w\|=1$ 的约束了。然而这个时候目标函数仍然不是凸函数，没法直接代入优化软件里计算。我们还要改写。前面说到同时扩大 w 和 b 对结果没有影响，但我们最后要求的仍然是 w 和 b 的确定值，不是他们的一组倍数值，因此，我们需要对 $\hat{\gamma}$ 做一些限制，以保证我们的解是唯一的。这里为了简便我们取 $\hat{\gamma}=1$。这样的意义是将全局的函数间隔定义为 1，也即是将离超平面最近的点的距离定义为 $\frac{1}{\|w\|}$。由于求 $\frac{1}{\|w\|}$ 的最大值相当于求 $\frac{1}{2}\|w\|^2$ 的最小值，因此改写后结果为：

$$\begin{aligned}&max_{\gamma,w,b}\ \frac{1}{2}\|w\|^2\\&\text{s.t.}\quad y^{(i)}(w^Tx^{(i)}+b)\geqslant1,\ i=1,\cdots,m\end{aligned}$$

这下好了，只有线性约束了，而且是个典型的二次规划问题（目标函数是自变量的二次函数）。代入优化软件可解。

到这里发现，虽然没有像其他方法一样先画好图，画好分类超平面，在图上标示出间隔那么直观，但每一步推导有理有据，依靠思路的流畅性来推导出目标函数和约束。

2.2.3　支持向量机的应用案例

下面我们用一个实例来讲解支持向量机（SVM）在乳腺癌影像诊断中的应用。

1. 案例

本案例的工作中，我们应当在基于乳腺癌数据库的数据挖掘和知识发现方面作出更多示范，为病人早诊断早治疗提供更多帮助。为此，本案例就以下几方面的进行研究与探讨：

① 研究如何对乳腺癌 CT 图像进行更好的预处理，为自动提取特征做出有效的准备，选择准确、有效的分类算法，最终获得满意的数据挖掘结果。

② 探讨和开发适合乳腺癌影像数据库知识发现的一般方法和工具。

③ 开发出性能良好的、易于医生操作的、接近医学专家水平的具有临床实践应用价值的乳腺癌辅助诊断系统。

2. 方法

本案例来源于一个在实际应用中发现问题、提出问题和解决问题的过程。本案例提出了将医学图像的处理技术与数据挖掘方法有机地结合，构造进行医学图像数据分类器的学习机制，通过从大量的乳腺癌 CT 图像数据挖掘出有用的信息，研究医学图像数据的纹理、形状特征提取技术和支持向量机的模型分类技术，从而有效地帮助医生准确地找出病灶区和疾病的程度，辅助医生进行诊断检查，提高诊断准确度。因此，该项目的研究具备重要的理论意义与广阔的应用前景。

首先介绍了图像处理的一般思想，给出了图像去噪、增强和分割的经典处理方法。利用边界保持类滤波器中的 KNN 平滑滤波器进行图形去噪，给出了去噪后的效果图。详细介绍了经典的直方图均衡化增强技术，给出通过这种技术对图像进行增强后的效果图。采用了基于区域增长的图像分割算法对乳腺癌图片的感兴趣区域进行分割，有效地分割出了潜在的肿块区域，为提取肿瘤形状特征奠定了基础。乳腺癌 CT 图片经过预处理之后，根据乳腺图像的特点，实现了基于图像边缘形状特征的提取方法，用于 MPSVM 分类器的输入，从而实现病例良性、恶性的判定。

（1）图像预处理

乳腺癌 CT 图像预处理的第一步是图像去噪处理，接着是图像增强处理。本文研究图像去噪、图像增强的相关算法。

（2）特征选取与提取

特征选择是非常重要的，其结果直接影响着分类器的性能。本文通过分析乳腺癌 CT 图像的特点，实现了基于灰度共生矩阵的纹理特征和灰度统计特征提取方法，基于形状特征提取方法。由于纹理特征中只有少数特征具有决定性的意义，因此我们应找出特定性特征并提出简单的规则，以快速地确定病人的病情。本文采用将约简后的属性集作为分类系统挖掘的输入，这可以减少分类器的输入的维数。

（3）改进的分类算法设计

本案例分析了标准 SVM 算法的不足之处，实现了将近似支持向量机分类算法应用乳腺肿瘤的分类检测中，并且指出了核函数在分类器设计中的重要性。实验表明：径基核函数的性能是最好的；

PSVM 算法比 SVM 算法速度快、对硬件资源要求低、易实现并且效果理想。

在我们实际应用 PSVM 算法进行乳腺肿瘤分类时，发现分类器过度拟合样本量的那一类数据，样本量小的数据分类准确率低。为了解决非平衡数据集的问题，本案例实现了 MPSVM 算法，该算法能消除两类样本点数差对整体分类性能上的影响。

（4）传统方法与本案例方法的区别

传统的分类方法存在一个比较严重的缺陷，它需要在样本数目很大的前提下进行分类，只有当样本数非常大时才能保证分类有较高的准确率，因此，它不适用于乳腺癌 CT 图像的分类检测。由于 PSVM 对于小样本分类空间同样具有较好的适用性，所以，对分类器进行设计只需要数目较少的样本就能够快速有效地训练出性能较好的分类器。

乳腺癌 CT 图像数据产生的 n 个特征向量，需要经过计算支持向量和二次规划后才能完成一次训练。而近似支持向量机不需要计算这些，只需求解一个线性方程组，所以速度非常快。因此，PSVM 算法对于乳腺癌 CT 图像分类中高维特征空间的情况特别有用。

对乳腺癌 CT 图像数据进行正常/异常、良性/恶性分类检测，PSVM 分类器实现过程如下所述：

① 对乳腺癌图像的基本预处理包括图像去噪、图像增强、图像感兴趣区域分割等。

② 提取特征向量，将得到的 n 个特征值作为分类器的输入。

3. 原理

在乳腺癌辅助诊断系统中，我们使用的原始 CT 图片包含大量有噪声的背景，在图像进行传输过程当中又会遇到噪声污染的情况发生，使得图像看起来有的太亮，有的又太暗，不能达到我们预想的效果。经过去噪声处理后，可以去掉图像中的大多数噪声、背景信息和孤立点。在需要保持原始 CT 图片信息的基础上，就需要抑制噪声，图像去噪技术就得设计合适的噪声抑制滤波器。常见的有三种滤波器：中值滤波器、均值滤波器和边界保持类滤波器。

在数据预处理中，使用边界保持类滤波器中的 K 近邻（KNN）平滑滤波器。KNN 平滑滤波器实现算法。

① 制作一个以待处理的像素为中心的 $m \times m$ 模板。

② 在该模板中能选择与待处理像素的灰度差为最小的 K 个像素。

③ 用 K 个像素的灰度平均值替换所有待处理的像素值。

如图 2-7 所示的 3×3 模板，K=4。对于像素 I（4，3）的值为 4，其相邻灰度差最小的像素点集为（4，5，5，5），可求得它们的平均值为（4+5+5+5）/4=4. 75，约等于 5。所以可得到 I（3，2）的值应为 5。

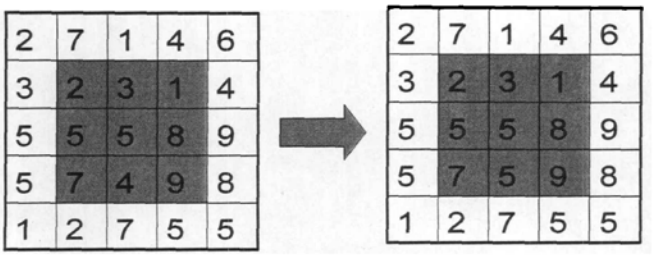


图 2-7　像素的灰度特征提取图

资料来源：高妮. 支持向量机及其在乳腺癌辅助诊断系统中的应用研究. 西北大学硕士论文.

图像增强技术的主要目的在于对图像进行加工处理，以得到大量有用的并且是可以使用的图像。图像增强可以突出图像中的某些信息，并且又削弱某些不必要的信息，达到很好的显示效果，使图像细节更清晰又易于辨认，得到的结果图像更适合人的视觉特性或者机器识别系统。对病灶区域边缘锐化并提高病灶区域和背景对比度，这两方面都是乳腺癌 CT 图像增强应达到的主要目的。但是我们对图像的平滑区域和边缘区域做灰度改变，会遇到图像的边缘部分还是不明显的情况，并且平滑区域也会丢失一些信息，不能得到我们想要的结果。

针对乳腺癌 CT 图像的相关特点，本文叙述直方图均衡化算法应用到医学领域的方法。

直方图均衡化是一种简单而易用的传统图像对比度增强算法，将原始图像比较集中的某个灰度区间变成全部灰度范围内的近似均匀分布，这是直方图均衡化的主要思想。效果如图 2-8 和图 2-9 所示。

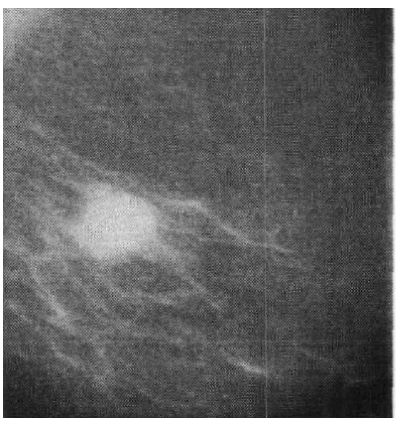


图 2-8　乳腺 CT 原始图

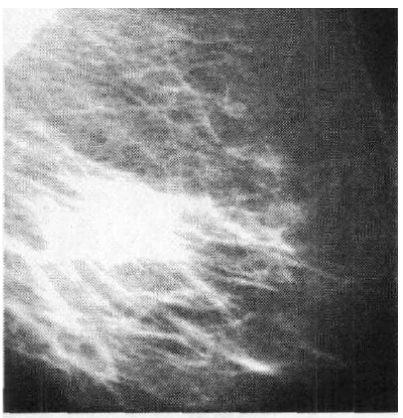


图 2-9　乳腺 CT 增强图

该算法通过增强函数 $t = EH(s)$ 来对图像空间域上的点进行增强，设目标图像和原始图像上的像素点 (x, y) 分别是 t，s。

在进行直方图均衡化增强过程当中，有以下两个条件可以满足增强函数 EH 的需要。

① 在没有打乱原始图像的灰度排序次序时，在 $0<-s<-M-1$ 的范围内，增强函数 $EH(s)$ 被认为是一个单调的并且递增的函数。

② 为了保证在变换的过程当中灰度值的动态范围是一致的，在 $0<-s<-M-1$ 范围内应该有 $0 < EH(s) < -M-1$ 成立。完成 s 到 t 的均匀分布转换可以通过增强函数来实现，从而得到增强转换方程为：$th = EH(sk) = \Sigma(ni/n) = \Sigma ps(si)$，($k$=0，1，…，$M$–1)。

在实际处理问题当中，文献给出了如下的步骤：先统计原始图像的灰度，计算原始直方图分布，进而可以再计算出累计直方图分布 tk，源灰度、k 到 tk 的灰度映射关系可以按照 $tk=[(N-1)xtk+0.5]$ 取整得到，其中灰度级数为 N。

对上面所陈述的步骤进行重复，便能获得所需的全部的源图像各灰度级到目标图像各灰度级的映射关系，然后，对源图像各点像素作灰度转换操作，这样就实现了源图像的直方图均衡化增强。

乳腺癌 CT 图像是乳腺癌病变检测模块的数据来源，在系统中它构成了基于纹理特征的 MPSVM 算法的第一分类器的输入数据，而第一分类器在本系统中处于正常/异常、良性/恶性检测模块中三个分类器的最前端，若该 CT 图像数据经第一分类器判定为正常，则检测处理流程结束，若被判定为异常则将该 CT 图像数据作为基于纹理特征的 RS-MPSVM 算法的第二分类器的输入。类似地，如果该 CT 图像被判定正常则流程结束，若为异常则利用基于形状特征的 MPSVM 算法的第第三分类器进一步对该 CT 图像进行处理。由第三分类器输出处理结果为良性或者恶性。

本乳腺图像的读取。图像预处理模块主要功能实现对读取的模块按次序进行相关处理，包括基于 KNN 平滑滤波器技术的图像去噪和基于直方图均衡化增强和粗糙集增强技术的图像增强。图像分类模块的功能是对经过预处理的图像按第一、第二和第三分类器的顺序进行分类处理，该模块包括纹理特征提取、特征离散化与约简、肿瘤区域分割和形状特征提取等重要的操作，在这些操作中几乎应用了前面几章所提及的相关算法。系统应用逻辑模块主要负责处理系统在该应用中处理流程、逻辑的控制问题。显示模块主要是指所处理图像的显示与更新问题。图像及训练数据保存模块主要负责处理后的图像和训练数据的保存，将其保存入医学影像数据库中。

表 2-9 径向基核函数参数 6 对分类的影响

核参数 6	运行时间（S）	正确率	支持向量机个数
100	85.3207	82%	2800
10	95.2684	90%	2800
1	450.9966	94%	1800
0.1	1997.5532	97%	980
0.01	453.2534	97%	2100
0.001	102.9930	97%	2800

资料来源：高妮. 支持向量机及其在乳腺癌辅助诊断系统中的应用研究. 西北大学硕士论文.

4. 意义

如表 2-9 所示，本案例的数据挖掘意义是巨大的，即使是图像的人工智能处理也可以用“支持向量机”这样一个算法来解决。首先搜集乳腺癌的 CT 图像数据，剔除噪音，用算法来提取图像纹理与灰度特征，分别建立良性/恶性识别模块，其中的图像增强，灰度特征提取，纹理特征提取都是数据挖掘算法完美的展现，把一帧图像划分为无数个点阵与方格，依据每一个点及其邻近点的深浅特征比对来识别正常图像与异常图像，这就是算法的力量，把困难的问题简单化。

2.3 疾病规律与统计学革命

2.3.1 肝胆外科的统计学故事

疾病规律与统计学有先天的联系，这是因为科学的推理与演绎、归纳与总结都离不开大样本的数据支持。只有在大样本的数据支持下作为个案的成功才能被推广为疾病的规律。可推理、可重复、可实验是科学的三大必然要素，其中数据是最重要的证据之一，肝胆外科中 ALPPS 手术的故事充分说明了这一点。

ALPPS 手术的数据挖掘故事

2007 年德国医生 Schlit 在一次手术中偶然发现：沿镰状韧带离断肝脏，同时结扎右门静脉可以在七天内促使左肝脏急速的显著增生（74%～87%）。这个意外的发现使得外科医生 Schlit 敏感地认识到，充分利用肝脏生长的这一特性也许意味着手术技术的一个新飞跃。很多肝癌晚期患者以往由于肝脏切除比例过大最后因肝脏衰竭而死亡，如果在离断肝脏的同时结扎右门静脉，左肝在一周内就会急剧生长，这就为切除病灶带来了更多的余量，肝癌晚期患者得到救治的可能性大大提高。基于这一原理，2011 年～2012 年德国医生报道了 5 例 ALPPS 手术。在缺乏大样本数据的支持下，ALPPS 手术作为肝胆外科的革命性技术迅速传入中国，全国各地的医院纷纷开展这个手术，抢先报道 ALPPS 手术的成果，在远期结果还远远没有显露的情形下 ALPPS 手术成为肝胆外科的潮流。

如表 2-10 所示，这一“革命性”手术技术的快速传播脱离了科学的规律，脱离了大样本数据的支撑，马上带来了随之而来的恶果，围术期的高死亡率，围术期的腹腔粘连、感染、胆漏、肝功能衰竭频频发生。ALPPS 手术本来是一个比较大的手术，需要各国医生国际协同，在小样本的基础上大胆革新，小心总结，在形成指南的条件下再慢慢推广。然而，ALPPS 手术的超常规扩散既为患者带来了福音，也产生了一定的负面作用。

表 2-10　近 3 年 ALPPS 手术在各国应用文献报道情况表

作者（出版年）	国家	例数	手术间隔天数	肝衰竭发生率
Schnitzbauer（2012）	德国	25	8	-
Li（2013）	德国	9	13	22%
Knoefel（2013）	德国	7	6	-

续表

作者（出版年）	国家	例数	手术间隔天数	肝衰竭发生率
Alvarez（2013）	阿根廷	15	7	20%
Sala（2013）	阿根廷	10	7	20%
Torres（2013）	巴西	39	14	2.5%
Dokmak（2012）	法国	8	7	-
Schadde（2014）	瑞士	48	-	12.5%

本故事中我们看到了医学实践与数据挖掘先天的联盟关系，少数的和个体的医疗行为能否成为一种可以推广的模式，大样本的数据提供了“不可辩驳的事实”，这就是临床医学的归纳、总结与推理，遵循严格的科学范式。

尽管 ALPPS 手术近年来存在 20%的肝衰发生率，它的产生与发展毕竟是肝胆外科手术的革命性进展，突破了中晚期肝癌患者的手术禁区。沿镰状韧带的肝离断与右门静脉结扎是 ALPPS 手术中最具革命性的创意，它虽然发现于偶然之中，却是对传统的“二步肝切法”的巨大改进，诱发左肝 87%的 7 天增生为千千万万的晚期肝癌患者的手术切除带来了可能，不能不看到，这是肝胆外科技术的巨大进步。

面对 ALPPS 手术的缺陷，数据技术与临床医学的结合有可能总结且观察出更好的解决方案。比如有的数据已经显露出腔镜手术比开放手术在 ALPPS 术式中有更好的预后，胆漏、粘连、感染的发生率都大大降低。从这个案例中我们也可以发现：所谓数据技术，其本质是对临床记录的一种知识发现。

2.3.2 双盲实验的诞生

双盲试验，是指在试验过程中，测验者与被测验者都不知道被测者所属的组别（实验组或对照组），分析者在分析资料时，通常也不知道正在分析的资料属于哪一组。该词通常用于医学领域。在药物测试中经常使用双盲测试。病人被随机编入对照组及实验组。对照组被给予安慰剂，而实验组给予真正药物。无论是病人或观察病人的实验人员都不知道谁得到真正的药物，直至研究结束为止。不过部分的试验会较难做成双盲，例如：如果治疗效果非常显著，或治疗的副作用非常明显，实验人员便可能猜想到哪组是对照。

案例：

2003 年，德国医生 Leucht 在著名医学杂志 Lancet 上发表了一篇文章，为 31 项共 2320 例抗精神病药双盲试验作了荟萃分析（meta-analysis ），说明新一代抗精神病药不一定优于老药。2008 年，他在世界著名的美国临床精神药理学家戴维斯（J M. Davis）教授的指导下，作为美国国立精神卫生研究所（NMH）的课题，做了一项大规模的双盲试验荟萃分析，这项研究包括 150 项合乎标准的双盲试验，共计 21533 例，数据记录如表 2-11 和表 2-12 所示。

他们的结论是：原先认为第二代抗精神病药（以前称为非典型抗精神病药）的特点是：①对精神分裂症的疗效较好，尤其是对阴性症状；②锥体外系症状（EPS）副反应较少。但是，从研究结果看来并非如此。

① 从疗效来看，只有氯氮平、氨磺必利、奥氮平和利培酮 4 种第二代抗精神病药，比第一代抗精神病药（以前称为经典抗精神病药，如氟派定醇、氯丙嗪等）稍好。

表 2-11　第二代抗精神病与第一代抗精神病药（氟哌啶醇 12mg/d 或氯丙嗪 600mg/d）双盲对照试验（包括药厂赞助项目）的荟萃分析的效应值（Hedge'sG）

项　　目	氯氮平	氨磺必利	奥氮平	利培酮	阿立哌唑	喹硫平	齐拉西酮
总疗效	-0.52**	-0.31**	-0.28**	-0.13**	-0.05	0.04	0.04
阳性症状	-0.36**	-0.22**	-0.15**	-0.13**	0.03	0.14	0.03
阴性症状	-0.27**	-0.27**	-0.32**	-0.13**	-0.09	0	-0.09
抑郁症状	-0.51**	-0.37**	-0.27**	-0.10	-0.12*	-0.23*	0.01

注：*P<0.05**P<0.01

表 2-12　第二代抗精神病与第一代抗精神病药（氟哌啶醇 12mg/d 或氯丙嗪 600mg/d）双盲对照试验（剔除药厂赞助项目的结果）的荟萃分析的效应值（Hedge'sG）

项　　目	氨磺必利	奥氮平	氯氮平	利培酮	阿立哌唑	喹硫平	齐拉西酮
总疗效	-0.31**	-0.28**	-0.22**	-0.00**	-0.05	0.04	0.04

注：*P<0.05**P<0.01

资料来源：颜文伟. 上海交通大学附属医学院精神卫生中心. 2009 年.

② 从整组看来，对于阴性症状没有突出优势。实际上，如果某药疗效较好的话，对阳性、阴性症状都有较好效果；疗效较差的话，对阳性、阴性症状都较差。

双盲试验是实验心理学中一个很好的控制额外变量的方法，是排除法的一种。双盲控制时让实验的操作者和实验被试都不知道实验的内容和目的，由于实验者和研究参加者都不知道哪些被试接受哪种实验条件，从而避免了主、被试双方因为主观期望所引发的额外变量。现代医学的科学性首先奠基于生物学和生理学之上。不过即便是最顽固的还原论者也会承认，医学毕竟不等于生理学，除了研究生理现象之外，医学的基本主题终究是“治病救人”。那么医学作为治病救人之学，在生理学之外，还有没有科学性呢？当然仍是有的。而最能体现医学作为“治病之科学”特性的，恐怕要算“双盲实验”的运用了。盲法或双盲法的思路很早就提出了，不过广泛应用于医学领域，大概是 20 世纪的事，因此有人把双盲法称作“20 世纪重大的科学进步”。双盲法被用于检测药品或治疗手段的效果，但其意义绝不限于药学，乃至成为医学的现代化或科学性的标志和象征。人们认为：“双盲研究已经引起了一场医学革命……医学必须建立在双盲研究的基础上，这个提法已经被理解为一场‘基于证据的医学’运动。”现代医学用双盲法规范和标榜自己，而当西方攻击中医时，双盲法也成为最常用的武器——有多少中药通过了双盲法的严格检验？的确，即便所宣称的“迄今还没有任

何中国的草药方剂得到了双盲研究的确证”恐怕是过于夸大，但可以相信，中草药面对双盲测试时确实成绩不好。

双盲试验引发了医学的思维与方法革命，直接导致了循证医学时代的到来。

小知识：荟萃分析

（1）荟萃分析概念

荟萃分析的概念最早是由 Light 和 Smith 于 1971 年提出的。当时针对大量发表的科学论文中，对于同样的研究却得出截然不同结果的问题，他们提出应该在全世界范围内收集对某一疾病各种疗法的小样本、单个临床试验的结果，对其进行系统评价和统计分析，将尽可能真实的科学结论及时提供给社会和临床医师，以促进推广真正有效的治疗手段，摈弃尚无依据的无效的甚至是有害的方法。

1976 年 Glass 首次将这一概念命名为 Meta-analysis（荟萃分析），并定义为一种对不同研究结果进行收集、合并及统计分析的方法。这种方法逐渐发展成为一门新兴学科——“循证医学”的主要内容和研究手段。荟萃分析的主要目的是将以往的研究结果更为客观地综合反映出来。研究者并不进行原始的研究，而是将研究已获得的结果进行综合分析。

（2）荟萃分析的分类

通常概念下的文献综述是对有关文献的内容或结果进行罗列、简单的描述和初步的讨论，而荟萃分析则完全上了一个台阶。根据荟萃分析所依据的基础或数据来源可以将其分为三类：文献结果荟萃分析（Meta-analysis based on literature, MAL）；综合或合并数据荟萃分析（Meta-analysis based on summary data, MAS）；独立研究原始数据荟萃分析（Meta-analysis based on individual patient data, MAP or IPD Meta-analysis）。它们的区别在于：MAL 的文献检索局限于已经发表的研究，然后将这些研究的结果合并进行分析；MAS 不仅要得到相关的发表的文献，同时还有作者进行的相关统计学数据的总结；而 IPD 荟萃分析除了要检索所有已发表的相关文献，还要寻找存在于各科学团体中的未发表的有关研究，在 MAS 基础上更进了一步。所有临床试验不管是否已经发表，必须能够从研究者处得到单个患者原始的，以及各效应指标的数据。这一点对于肿瘤病因或疗效研究方面的分析来说较为重要。因为多数的关于肿瘤病人预后的 III 期临床试验，主要的研究指标大多为生存时间或生存率，或疾病无进展时间等，在多数情况下，不同的出版物中所得到的信息不足以进行一项真正的事件（如肿瘤死亡）发生时间全过程的分析。这使得以已经发表的文献作为基础的 MAL 和 MAS 变得较为困难。同时，考虑到有统计学意义的阳性结果较阴性结果更易发表等能够造成偏倚发生的情况存在，故 MAL 和 MAS 有一定的不足。相对来讲，IPD 荟萃分析不存在上述的弊端或受有关偏倚的影响较小。因此，在肿瘤生存或疗效研究领域中，当要求进行这方面的分析时，IPD 荟萃分析是唯一推荐使用的分析方法，尽管它比其他两种方法要耗费更长的时间，以及人力和物力。

2.3.3　几则很有趣的医学统计学故事

医学统计学是一门很奇妙的科学。要说它简单吧，其实也挺简单的，常见的统计方法也就十余种，在教科书上都能找到，只要熟练掌握，虽不敢夸下海口说可以“以秋风扫落叶的气概横扫四海之内的杂志”，但足以轻车熟路地应付 99%的科学研究。要说它复杂吧，也挺复杂的，毫不夸张地说，绝大部分国内期刊，甚至在很多低分 SCI 杂志上，乱用统计学的现象多如牛毛。

很多同行在学习医学统计学时，都在抱怨自己很难走出“一学就会，一会就用，一用就错，一错就懵”的怪圈。究其原因，主要是部分同行学习医学统计学时都抱着一副“依葫芦画瓢”的态度，试图“套用统计学方法”来解决自己面临的问题，而不去仔细思考统计学方法的来龙去脉。本文拟谈几则与医学统计学相关的故事，希望能帮助大家从宏观上正确认识医学统计学这门科学。

1. 两个指标诊断疾病的问题

（1）肝癌诊断指标的优劣

路人甲做了一个研究，旨在比较两个指标（A 和 B）对肝癌的诊断价值。路人甲以 A 和 B 的参考范围上限作为诊断界值，得出了 A 和 B 在该界值下对应的诊断敏感性和特异性。结果表明：A 的诊断敏感性为 0.80，特异性为 0.90；B 的诊断敏感性为 0.85，特异性为 0.87。路人甲很快撰写论文报道了自己的研究成果，指出 B 诊断肝癌的敏感性高于 A，而特异性低于 A。

路人乙是这篇文章的审稿人，当他看见这个结论后，脸色铁青，毫不犹豫地在审稿意见中写道：就敏感性而言，B 高于 A；就特异性而言，A 高于 B。诊断敏感性和特异性与所采用的界值密切相关，作者得出的敏感性和特异性仅仅代表了一个诊断界点下面的诊断效能，无法从全局上反映 A 和 B 的诊断价值。文章的结论到底是想说明 A 优秀还是 B 优秀呢?

这个故事说明：统计指标选错了，统计出来的东西往往难以“自圆其说”。

稿件被退了，路人甲有些许郁闷。经过认真学习科研设计与统计学知识后，路人甲终于明白了一个问题：两个指标诊断性能的比较是不能比较敏感性和特异性的，而应该比较 ROC 的曲线下面积，因为曲线下面积才是衡量整体诊断效率的最佳指标。路人甲很快绘制了 ROC 曲线，统计结果表明，A 的曲线下面积为 0.80，B 的曲线下面积为 0.82。路人甲欣喜若狂，赶紧动笔写论文，并且理直气壮地给文章定了一个结论：B 的诊断效率是优于 A 的，其理由就是因为 B 的曲线下面积大于 A。

路人丙是这篇文章的审稿人，当他看见这个结论后，脸色铁青，毫不犹豫地在审稿意见中写道：从表面上看，B 的曲线下面积高于 A，但是导致这种差异的原因有两种，一种是抽样误差，一种是试验效应，即 B 确实是高于 A 的。你怎么能确定这不是抽样误差呢? 在统计学上，要确定 0.82 是否高于 0.80，就一定要经过统计学检验的。

这个故事说明：在医学科研中，没有经过统计学检验的结论多半是不科学的。

稿件被退了，路人甲很是郁闷。他吸取了经验教训，自学了很多统计学理论，终于弄清楚了采用何种方法去比较曲线下面积。接下来的事情就是改稿，然后另选杂志继续投稿。路人甲在文稿中特别注明了，曲线下面积是经过了统计学检验的，B 的曲线下面积（0.82）与 A 的曲线下面积（0.80）之间的差异是有统计学意义的，而且还大摇大摆地在后面加了个括号，写明 P=0.01。路人甲仰天长

叹了一口气，很郑重地给自己的研究下了结论：本研究表明 B 的诊断效率是优于 A 的。

路人丁是这篇文章的审稿人，当他看见这个结论后，脸色铁青，毫不犹豫地在审稿意见写道：B 是常见的诊断指标，其检测结果并不对临床医师设盲，在很大程度上可以影响临床医师对疾病的诊断。A 是新近发现的诊断指标，其结果完全对临床医师设盲，不可能影响医生的诊断。所以作者的结论（B 比 A 优秀）是不可靠的。

再说得通俗点，如果把 A 和 B 分别理解成法庭上的原告和被告，那 B 无疑既充当了辩护律师，又充当了法官的角色。在这种情况下，A 输掉了官司是十分正常的。如果换一个公平的、独立的法官来断案，B 能否胜出就不好说了。

这个故事说明：实（试）验设计有缺陷，再优秀、再正规的统计学方法也于事无补。

稿件又被退了，路人甲的心情极度郁闷。思来想去，决定把实验重做一遍，让 A 和 B 在一个公平的环境中比较（为便于描述，此处忽略医学伦理学问题）。在新开展的研究中，A 和 B 都是对临床医生设盲的，不可能影响诊断标准。这下 A 和 B 的比较结果应该比较可靠了吧，路人甲又仰天长叹了一口气，感觉自己如释重负了。

科研太折腾人了，太不容易了！统计结果很快出来，A 的曲线下面积是 0.80，B 的曲线下面积则变成了 0.77，经过统计学检验后发现，A 的诊断效能确实是高于 B 的。整个研究的试验设计滴水不漏，统计学过程天衣无缝，我就不信还有人敢拒这篇稿件，路人甲心中开始暗喜。

路人戊是这篇文章的审稿人，当他看见这个结论后，脸色铁青，毫不犹豫地在审稿意见写道：A 和 B 的检测并不矛盾，他们之间的关系不应该是竞争关系，而应该是合作关系。读者最关心的问题显然不是 A 和 B“孰强孰弱”的问题，虽然这个问题有一定的专业价值。

如果我是坐诊医生，我会说：A 和 B 谁强谁弱关我什么事？总之来一个病人我就 A 和 B 都检测，我的患者都不差钱！作者的研究重点应该是明确 A 和 B 能否互补，联合使用是否能有助于提高诊断准确性的问题，而非 A 和 B“孰强孰弱”的问题。简单地说，就是明确 1+1 是否大于 1 的问题。

文章又被拒稿了。

这个故事说明：研究方向错了，即使是无懈可击的实（试）验设计和天衣无缝的统计方法，也是无济于事。

这四个故事说明：医学科研是很痛苦的，不重视统计学和科研设计，会走很多弯路的。

（2）降糖药的研究、学生自杀事件

路人甲长期从事降糖药的研究，最近他发现了一种药物，可以降低患者的血糖。为了评价该药的降糖效果，路人甲费尽心机地设计了一个看似完美的随机对照试验（RCT），为了保证结果可靠，路人甲严格遵守 RCT 设计准则，包括随机、双盲、安慰剂对照等措施。

研究结果表明，实验组和对照组在接受药物治疗前血糖浓度的均值都是 10mmol/L，差异无统计学意义，表明两组研究对象的基线特征具有可比性。对照组未经任何药物治疗（为便于描述，此处忽略医学伦理学问题），血糖浓度还是 10mmol/L；实验组经过药物干预后，血糖浓度变成了 9mmol/L。

统计学检验结果表明，实验组和对照组治疗后的血糖浓度的差异是有统计学意义的（$P<0.01$）。路人甲赶紧撰写论文，并毫不客气地给研究下了个结论：该药可以降低患者血糖。

路人乙是这篇文章的审稿人，当他看见这个结论后，脸色铁青，毫不犹豫地在审稿意见写道：该药确实可以降低血糖，但是一个只能降低 1mmol/L 的降糖药有何临床价值?

这个故事说明：有统计学意义不一定有专业意义。

路人甲做了一个调查，同处一地的 A 和 B 两所中学，各有 1000 名学生，过去的一年，A 校有 5 名学生自杀（自杀率为 0.5%），B 校没有学生自杀（自杀率为 0%）。统计学结果表明，两校自杀率的差异无统计学意义（P=0.07，Fisher 确切概率法，笔者进行了统计）。于是路人甲得出结论：A 和 B 两校的自杀率是没有差异的，A 校 5 名学生自杀纯属小概率事件。

路人乙是这篇文章的审稿人，当他看见这个结论后，脸色铁青，毫不犹豫地在审稿意见写道：5 个鲜活的生命就这样没有了，5 个家庭就这样毁了，你却告诉我这纯属小概率事件，你就不怕“人神共愤”吗?

这个故事说明：有专业意义不一定有统计学意义。

这两个故事说明：做医学科研，不能死磕统计。

看完这两个故事，也许有的读者会有疑问：前面还强调“没有经过统计学检验的结论多半是不科学的”，为什么这里却淡化统计学的作用呢? 对此，笔者认为：统计学仅仅是一种工具，用得好当然可以事半功倍，但是在某些情况下，工具往往就是个累赘，也许徒手干活才是最好的选择。

2. 如何看待统计学结果

路人甲经历数十年的研究，动用了各种高精尖的研究手段，发现了一个新的蛋白（命名为蛋白 A）。在肝癌患者中展开的研究表明，蛋白 A 和甲胎蛋白（AFP）有很好的相关性，其相关程度之好，几乎可以用“一塌糊涂”来形容，相关系数达到了 0.99（P<0.0001）。路人甲欣喜若狂，尽管蛋白 A 的检测过程还十分繁琐，检测费用还十分高，但是路人甲还是把持不住内心的激动，日夜兼程地撰写论文，宣称自己找到了一个新的肝癌标志物。

路人乙是这篇文章的审稿人，当他看见这个结论后，脸色铁青，毫不犹豫地在审稿意见写道：统计结果表明蛋白 A 和 AFP 的相关性十分明显。如果是这样，在临床实践中，通过检测 AFP 完全就可以得知蛋白 A 的浓度了，蛋白 A 在肝癌中的临床价值完全可以被 AFP 代替，还不说蛋白 A 的检测过程繁琐，费用太高的问题，你说蛋白 A 还有什么价值?

这个故事说明：统计学阳性的结果未必是“好结果”。

路人甲发明了两套诊断肺癌的方案，分别命名为 A 和 B。为了明确这两种方案到底谁“更胜一筹”，路人甲找了 100 个肺癌患者和 100 个疑似肺癌患者（结核、肺炎等），分别用 A、B 两套方案去进行鉴别诊断。在 200 个研究对象中（100 个肺癌和 100 个非肺癌），方案 A 正确了 100 回，准确率 50%，方案 B 仅仅正确了 50 回，准确率仅为 25%。

卡方检验表明：方案 A 和 B 准确率之间的差异有统计学意义（P<0.01）。很明显，方案 A 的准确性要高于方案 B。路人甲赶紧发表论文，指出：方案 A 诊断肺癌的准确性优于方案 B。

路人乙是这篇文章的审稿人，当他看见这个结论后，脸色铁青，毫不犹豫地在审稿意见写道：如果我（审稿人本人）是坐诊医生，他就会反着看方案 B 的结果，凡是方案 B 认为是肺癌的，他就认为病人不是肺癌；反之亦然。这样下来，200 个病人中，方案 B 应该能正确识别 150 个人，准确率为 75%。

统计学结果表明，方案 B 的准确率（75%）是高于方案 A（50%）的，所以真实的情况是方案 B 优于方案 A。实际上，当面对这 200 名患者的时候，随便到城隍庙找个瞎子来“算命（猜患者是否患病）”，按照统计学理论，准确率也应该是 50%，方案 A 的价值可以说是“一无是处”。

这个故事说明：统计学阴性的结果未必是“坏结果”。

还是那个 AFP 与蛋白 A 的例子。路人甲发现蛋白 A 和甲胎蛋白（AFP）之间有很好的相关性，也开始撰写论文，但是他的结论并不是“蛋白 A 是诊断肝癌的标记物”。他认为，既然蛋白 A 与 AFP 之间有很强的相关性，那提示 AFP 和蛋白 A 之间可能存在十分密切的“调节”或者“被调节”关系。

最终，路人甲围绕“蛋白 A 与 AFP 之间的调节或者被调节关系”做了很多分子生物学试验，指出蛋白 A 是调节 AFP 表达的唯一（注意“唯一”这两个字）因子，因此二者之间才会呈现如此强烈的相关性。这是一项基础研究，虽然未能直截了当地地指出蛋白 A 的临床价值，但是这个研究形象生动地讲述了一个完整的分子生物学事件，丰富了我们对于肝癌发生与发展分子机制的认识。最终，论文“堂而皇之”地被接受了。

这个故事说明：同一统计学结果，从不同的专业角度去解释，结论是完全不同的。

这三个故事说明：对统计学结果的解读一定要结合专业！从专业中来，到专业中去！

资料来源：胡志德（Journal of Thoracic Disease 学术沙龙委员、Section Editor (Systematic Review and Meta-analysis)，工作于济南军区总医院实验诊断科，现为第二军医大学临床检验诊断学博士研究生，以第一作者或通讯作者身份发表 SCI 论文十余篇，并主持国家青年科学基金一项。）

2.4 老年肺癌研究

2.4.1 数据的抓取与来源

数据挖掘技术为医学研究开启了新的一扇窗口。通过互联网数据抓取技术对全球老龄肺癌进行研究的基本思路是抓取各种中文、英文网页中含有相关关键词的网页数据，下载到本地数据库中建立老龄肺癌的大型数据库。第二步是对这些数据进行格式化整理，包括分类、聚类、数据清洗，最后用语义分析的方法进行趋势研究，其中包括数据库技术、全文检索技术、统计学技术、数学模型与计算机算法统计技术，最终用技术图表的方式发布研究成果，这就是互联网数据挖掘技术在医学领域的最新应用成果。在文件检索与分析方面为临床医学提供了事半功倍的高效率工具。不仅仅如此，其中的定量定性分析方法、数理统计分析方法节省了临床医学所需要的大量药物与手术经验积累成本，直接导出新的研究成果为患者服务。

医学领域的数据构成一个复杂的数据库，包括电子病历、医学影像、病理参数、化验结果等，而目前数据挖掘技术主要应用于以结构化数据为主的关系数据库、事务数据库和数据仓库，对复杂类型数据的挖掘尚处在起步阶段。但是，随着数据库技术的发展，数据挖掘技术作为一种解决方案，成为医学信息技术领域重要的研究方法，必将为决策支持、科学研究带来很大的方便和可观的效益。

本文选取的主要研究对象就是全球老龄肺癌问题。通过对全球各种资料文献、卫生组织文件、

国内医学期刊发表的关键词文章、论文的语义分析，多维度研究老龄肺癌的分布特点、存活率，老龄肺癌的手术适用性、放化疗成功率等具有重大临床医学价值的关键点，为医务工作者提供一个强大的医学数据库与数据分析成果。如图 2-10 所示，我们选取 2000-2011 年期间发表的各国医疗文献中有关老龄肺癌的数据进行大规模下载与比对，先期建立本地数据库，用数理统计方法进行清洗、分类，最终用语义分析、回归预测等多项数据技术进行深度研究，我们把所有与老龄肺癌相关的数据在初筛中分为卫生组织文件、国外医学文献、国内医学期刊、文献四大类别。最高获取国内医学期刊八十九万页，最低获取国内医学文献二十三万页。

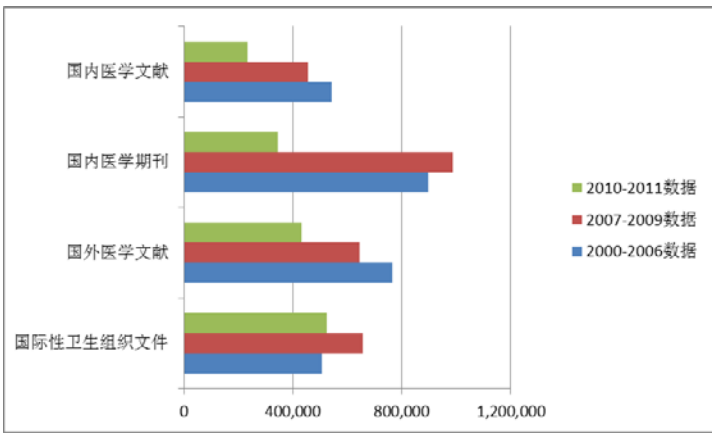


图 2-10　互联网医学文献抓取分类图

资料来源：邵学杰. 老年肺癌的最新统计学研究. 2012 年.

2.4.2　癌症与老龄化的相关性分析

癌症与老龄化的相关性分析最初来源于研究老龄手术的适应症与禁忌症，特别是需要考察老年人手术的约束条件。

图 2-11 是加拿大癌症与老龄化关系研究曲线。左纵坐标为每千人患病例数，右纵坐标为年龄，如图所示，年龄越大每千人患病数越多。国外医学上习惯以 70 岁以上老人为老龄人口，传统的人口统计学以 60 岁以上老人占总人口 7%的社会称为老龄社会。在从不同年龄人群肿瘤发病构成中，65 岁以上的老龄人群肿瘤发病率所占比重最大，约为每千人 60-90 例。分析表明，肿瘤发病率的上升取决于老年人群，老年人群肿瘤发病率的上升会导致全人群肿瘤发病率的上升。

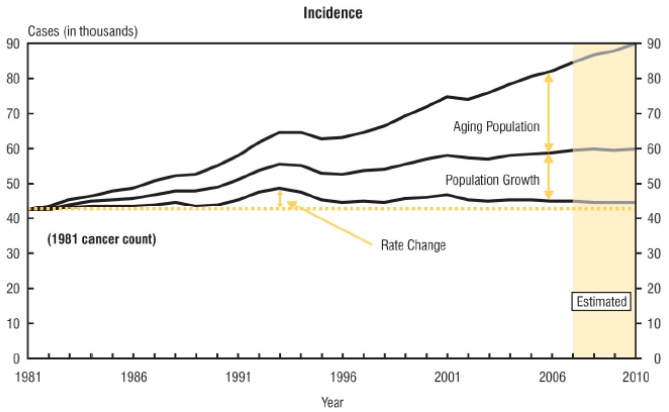


图 2-11　加拿大肺癌与老龄化关系曲线

资料来源：Canadian Cancer Statistics 2010

中国的情况也是如此。综合前两次调查的资料以及卫生部近年的报告，一条死亡曲线被勾画出来。

20 世纪 70 年代，我国每年死于癌症的人口约 70 万。城市癌症死亡率 91.8/10 万，占全部死亡人口 16.3%；农村癌症死亡率 80.8/10 万，占全部死亡人口 11.6%。

90 年代，我国每年死于癌症的人口约为 117 万。城市癌症死亡率 112.6/10 万，占全部死亡人口 20.6%；农村癌症死亡率 106.8/10 万，占全部死亡人口 17.1%。

21 世纪初，我国平均每年死于癌症的人口约为 150 万。城市癌症死亡率 124.6/10 万，占全部死亡人口 22.0%，在各类死因中居第 1 位；农村癌症死亡率 127.0/10 万，占全部死亡人口 21.0%，在各类死因中居首位。

而最新的数据来自 2006 年 5 月卫生部公布的《中国慢性病报告》——近年来癌症死亡人口已占我国城乡总死亡人口的 24%。

高发癌谱也发生了变化。《中国癌症控制策略研究报告》显示了 30 年来主要癌症死亡率排位的变化：

70 年代的排位是——胃癌、食管癌、肝癌、肺癌、宫颈癌。

90 年代的排位是——胃癌、肝癌、肺癌、食管癌、直肠癌。

2000 年的排位是——肺癌、肝癌、胃癌、食管癌、直肠癌。

其中，死亡率下降最明显的是宫颈癌，上升最明显的就是肺癌。

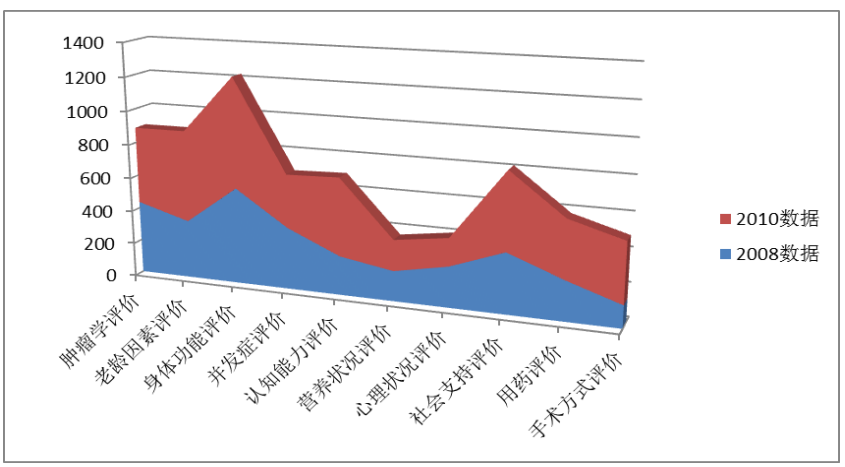
就全人群而言，肺癌是发病率最高的肿瘤，也是癌症死因之首，胃癌、食管癌和肝癌则紧随其后。男性 5 种最常见肿瘤依次为肺癌、胃癌、食管癌、肝癌和结直肠癌，占所有癌症病例的 2/3；女性最常见肿瘤依次为乳腺癌、肺癌、胃癌、结直肠癌和食管癌，占所有癌症病例的 60%，乳腺癌占所有女性癌症的 15%。值得注意的是，甲状腺癌在女性中呈现高发态势。2003 年～2011 年，女性甲状腺癌发病率年增长 20.1%。

2.4.3　老年人肺癌手术适用性评估关键词频率

如表 2-13 所示，老年人肺肿瘤切除手术的适用性，特别是 75 岁以上老人是否适合手术治疗仍然是社会与家庭很关心的问题。图 2-12 的 10 个指标构成了老年肺癌手术适用性的评价指标模型。

表 2-13　老年肺癌的手术评价要素

肿瘤学评价	一次标准的肿瘤学检查包括但不限于：彻底的病史及体检，支气管镜，CT/PET ＣＴ，血液检查，组织活检
老龄因素评价	身体功能，并发症探测，认知，心理，围术期护理，药物
身体功能	主要是与手术相关的心血管、肺功能测试
并发症	老年人过往病史研判，体检结果研判。能否预测哪些肺癌患者术后效果好、生存时间较长些？可用公用的预后评估模型察尔森合并症严重度指标（Charlson comorbidity index，CCI）为基础
认知能力	认知缺陷对术前检查很重要，对术后康复也有重大影响
营养状况	老龄人口手术风险主要是耐受性。老年人营养差，身体弱会对手术，化疗的耐受性产生重大影响
心理状况	30%的老年肿瘤患者都有心理疾病
手术方式选择	按照美国 SEER 数据库显示：右全肺切除要尽量避免，这与术后生存率息息相关。数据挖掘后还发现：年轻人肺段切和楔形切的生存率大大低于肺叶切除术，老年人则无此差别，就局部复发而言，肺段切高于肺叶切
社会支持因素	老人家庭的支持，围术期护理至关重要



单位：频次/50,000 字符

图 2-12　理念肺癌手术评价关键词频率图

一般来讲，老年人患恶性肿瘤的特点是肿瘤生长速度相对迟缓，病情发展也较慢，如果能得到有效治疗则复发的时间也较长，甚至可能不复发，很多患者最后死于其他疾病。对于肺癌患者来说，其主要的治疗方法有 3 种，首选的方法是手术治疗，其次是化学治疗（简称化疗）和放射治疗。其

他的还有分子靶向治疗、生物治疗等。如果病人选择得当，手术治疗可以根治肺癌。但遗憾的是约有 2/3 的肺癌病人因各种原因得不到手术治疗，这也是肺癌远期生存率低的重要原因之一。国际上公认的肺癌治疗原则是以手术为主的个体化多学科综合治疗，这当然也适合于老年人肺癌。

当然手术并非适合每一个人，尤其对于老年人来说，手术适应症的选择是很严格的。最基本的条件是要有良好的心脏和肺功能；没有其他系统的严重疾病；而且肺癌限于早中期病人；手术前要做好充分的准备，对病人存在的各种生理或病理的紊乱要予以纠正，尤其心肺功能有障碍的病人，要给予适当的支持治疗；术后针对病人的不同心肺情况，给予适合的治疗措施，必要时呼吸机辅助治疗等，以帮助病人度过术后的危险时期；病人完全康复后，应根据病人的不同情况，针对肿瘤给予不同的辅助治疗。

西医的传统观念认为手术预后与年龄有很大的相关性，由于术后并发症的死亡率高，老龄患者，特别是 75 岁以上老龄患者的手术治疗有严格的约束条件。然而，随着现代医学的发展，特别是小切口、电视胸腔镜（VATS）微创手术技术的进步，越来越多 80 岁以上高龄手术的成功使得很多的胸外科医生认为年龄不再是手术的禁忌。特别是 1 期、2 期的老龄患者医学界早已达成手术治疗与年轻人无异的预后共识。老年患者只要没有严重的阻塞性通气功能障碍，无近期心肌梗塞、无药物失控的每分钟 l0 次以上的室性早搏、无心力衰竭等，都可以做手术准备。

如表 2-14 所示，资料表明电视胸腔镜（VATS）技术在早期肺癌切除中并没有优势，这也是很多胸外科医生仍然偏爱小切口开胸手术的原因。老年肺癌手术从数理统计学看仍然有很多的并发症死亡率，虽然现代医学对老年患者的年龄敏感性降低，但手术前的 10 因素评估仍然是很重要的决策。近年来，随着腔镜技术的普及与发展，尽管预后与开放式手术相比在统计数据上没有太大的区别，但腔镜技术越来越普及。

表 2-14　胸腔镜与开放手术效果对比表

		Vats Lobectomy	Open Segmentomy	Significance
Mean%Predicted FEVI	(Range)	54(31-69)	51(34-69)	p=0.76
Mean Operating Time(mins)	(Range)	204(80-270)	195(114-266)	p=0.9
ITU admission		3(5.7%)	3(5.7%)	p=1
30 day mortality		3(5.7%)	3(5.7%)	p=1
Length of Stay(days)	(Range)	8.8(3-67)	10.4(3-32	p=0.97
Mean Survival	Years	6	5.4	p=0.98

资料来源：Congenital lung lesions and emphysema; lung cancer surgery

2.4.4　老年肺肿瘤的数据分析

老年肺癌高峰值区为 70~79 岁，如图 2-13 所示，无论巴西、加拿大、韩国均是如此。中国上海最高，平均每 10 万人发病 9~10 个，即十万分之九。有调查发现，六成的肺癌病者在初诊时已属后期，使得肺癌诊疗的整体结果仍不令人满意。老年性肺癌的发生几率越来越高，而且肺癌以鳞癌居

多，腺癌次之。一般老年肺癌的特点主要包括以下几点。

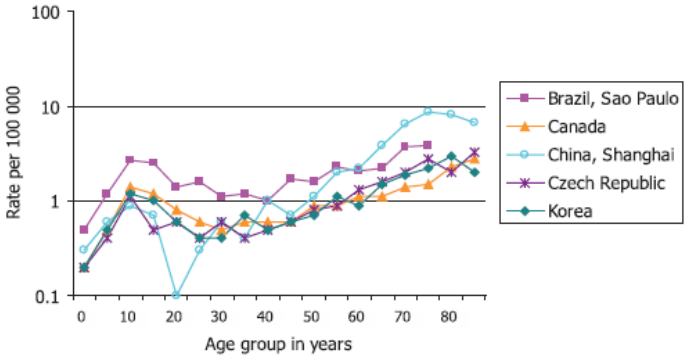


图 2-13　每百万人肺癌患病人数

资料来源：WHO report 2009-2010

一是起病缓慢，病程较长，患者病程平均 5.2 个月；二是临床表现以咳嗽、胸痛、痰中带血或少量咯血居多；三是 X 线胸片主要表现为块影，浸润性病灶、肺不张和胸腔积液；四是伴发病多，伴发的疾病主要是慢性阻塞性肺疾病和肺结核；五是误诊率较高，常被误诊为肺炎、肺结核和结核性胸膜炎；六是以男性居多，大多数有吸烟史。充分认识肺癌的特点，从医患两方面防止误漏诊，是我们做到早期发现、早期诊断、早期治疗的关键。

从人口统计学角度分析，欧洲地区人口年龄越大，相对生存率越低，这并不单纯是肺癌自身生存率低造成的，人口老龄化也是导致相对生存率低的重要统计学原因。换句话说，75～99 年龄组的老人即使没有肺癌，生存率也不会太高，但这并不能掩盖很多老人初诊时已是晚期的重要事实。这个统计学分析进一步导出了肺癌的特点：发现晚，发现难。特别是中国的肺癌发病率更是高于欧洲国家，这样的特性适合大规模的人口筛查。

如图 2-14、图 2-15 所示，肺癌与吸烟的关系早在 30 年前已经成为共识，吸烟因素能大大增加肺癌的几率。然而不吸烟的女性肺癌患者近年来增长很快，这凸显环境因素变化对肺癌的影响，我们将另文分析这一现象。

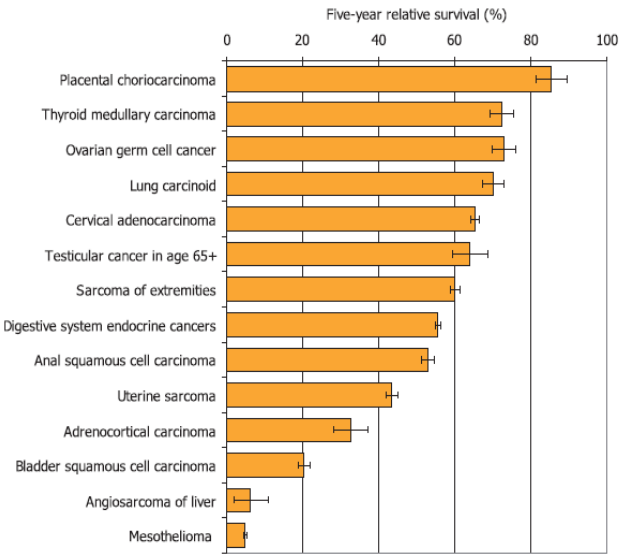


图 2-14　各种癌症的 5 年相对生存率

资料来源：WHO

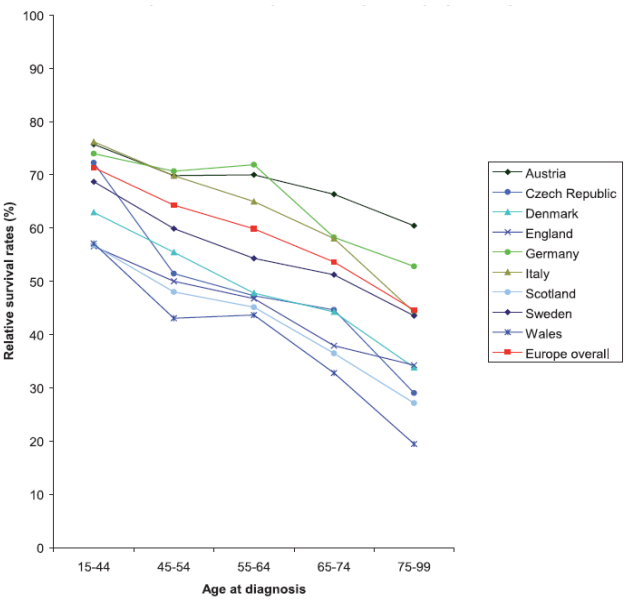


图 2-15　欧洲地区肺癌确诊年龄与相对生存率数据

资料来源：WHO

流行病学研究中常以 65 岁作为老年人群的界定标准，然而在临床试验中多以 70 岁作为老年患者筛选的上限，因 65～70 岁的患者被认为具有较好的健康状况，可以耐受适用于年轻患者的治疗方案并从中获益。而 70 岁之后机体各器官的功能明显下降，因而 70 岁被认为是机体衰老的年龄界限。既往对于老年 NSCLC 患者的判定标准多采用发达国家≤65 岁、发展中国家≤60 岁。而早在上世纪 70 年代，胸外科专家就曾专门针对老年肺癌进行过讨论,由于各种治疗方法的改进和有效的新药不断问世，增加了肺癌治疗的安全性；且肺癌多见于 60 岁以上，因此一致认为应以 70 岁以上作为老年肺癌的年龄标准，WHO 对此标准亦予以认可。检索且主要分布在 2006 年后发表的文章。显而易见，≤70 岁作为老年 NSCLC 患者的年龄界限已渐被公认且采纳。

资料来源：WTO

如图 2-16 所示，非洲地区肺癌只是排第十一位的癌症。这进一步说明肺癌与吸烟，空气污染等环境因素息息相关。其中男性患者是女性患者的 3 倍，死亡率中男性也高于女性 3 倍。性别差如此之大再次证明环境与人类行为是导致肺癌的重大因素。世界卫生组织下属国际癌症研究机构发布报告，首次指认大气污染对人类致癌，并视其为普遍和主要的环境致癌物。空气污染导致癌症是一个长期积累的过程，因此要确定目前肺癌高发与空气污染相关，需要追溯近 10 年的空气数据，但目前缺乏相关数据。

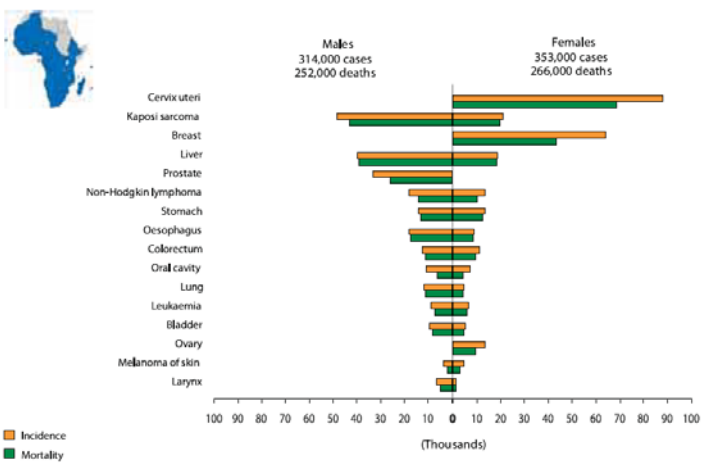


图 2-16　非洲地区肺癌分布特点数据

英国数据如表 2-15 所列。图 2-17 中左纵轴是英国每年肺癌患者死亡人数，右纵轴是英国肺癌患者每 100,000 人中死亡率，横轴是年龄组。数据统计显示了年龄与死亡率的深刻关系，75～79 岁的肺癌患者每十万人死亡率或年度死亡人数均处在高峰的年龄段，这表明男性 70～74 岁阶段，女性 75～80 岁阶段是肺癌筛查的最后警戒时间点。然而，现实中人类对老龄（70 岁以上）的肺癌筛查很不重视，一年两次以上的体检十分重要。无论男女，65 岁以后进入肺癌高发期，最佳预警时间是 60～64 岁，预警时间的前置对人类提高寿命，免于肺癌的打击至关重要。

表 2-15　英国肺癌患者年龄与生存率数据分析

Average Number of Deaths per Year and Age-Specific Mortality Rates per 100,000 Population, UK

Age Range	Male Deaths	Female Deaths	Male Rates	Female Rates
0 to 04	0	0	0	0
05 to 09	0	0	0	0
10 to 14	0	0	0	0
15 to 19	0	0	0	0
20 to 24	1	1	0	0
25 to 29	2	1	0.1	0
30 to 34	8	8	0.4	0.4
35 to 39	33	26	1.5	1.2
40 to 44	109	99	4.7	4.2
45 to 49	254	247	11.9	11.2
50 to 54	576	488	30.6	25.3
55 to 59	1,263	940	70.1	50.5
60 to 64	2,184	1,593	123.6	86.2
65 to 69	2,845	1,933	214.2	134.7
70 to 74	3,385	2,378	301.2	186.3
75 to 79	3,692	2,767	423.5	249.1
80 to 84	3,065	2,517	528	286.7
85+	2,329	2,180	552.9	238.8
All Ages	19,747	15,178	65.5	48.6

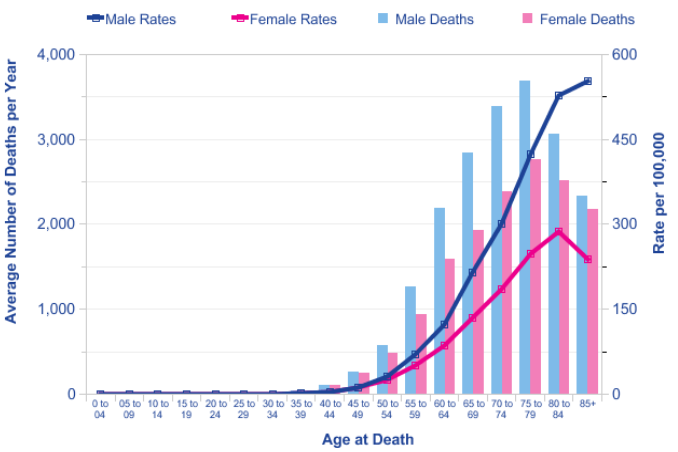


图 2-17　英国肺癌患者每十万人中的死亡率

在中年年龄组中，分析发现从 45～49 岁年龄组肺癌开始发轫，这表明从 40～44 岁开始进行肺癌专项体检十分重要。如果人类从 40 岁开始进行肺癌筛查，早期发现的存活率要高很多。

结论：从英国肺癌数据分析得到的启发是人类如果能够提早进入肺癌的筛查预警时间，存活率会大大提高。值得我们关注的肺癌发轫年龄为 45～49 岁，高峰死亡率 75～79 岁。最佳预警筛查时间为中年组 40～44 岁，至少一年一次体检。老年组 70～74 岁，至少一年两次体检。就老年人而言，要打破一年一次的体检惯例，70 岁以上老年人一年两次以上的体检是十分必要的重要肺癌预警方法。

吸烟与 23%～25%的国人癌症死亡相关，2010 年超过半数的成年男性是当前吸烟者，青少年男性中吸烟率还在攀升。即使这种吸烟率保持不涨，估计每年有 100 万例吸烟相关死亡，到 2030 年这一比例将翻倍。吸烟相关疾病将在吸烟二三十年后显现，即使推行控烟，接下来 10 年我国癌症负荷还会继续加重。我国癌症死亡例数从 2000 年的 51090 例增长到 2011 年的 88800 例，约 60%的癌症是可通过减少可控危险因素暴露来预防的。

2.4.5　英国肺癌患者 38 年来死亡率研究

如图 2-18 所示，从 1971 年～2009 年，英国每十万人中肺癌死亡率综合成缓降趋势，然而男性与女性有较大的差别，男性死亡率呈下降趋势而女性呈缓慢增长的趋势。女性肺癌患者 30 年来比较平稳的死亡率表明英国社会肺癌的环境因素在长达 30 年的过程中没有太大的变化，人口老龄化，职业女性，外来移民女性吸烟习惯的增加是重要的因素。男性肺癌患者死亡率的下降凸显现代医学进步带来的成果。吸烟与肺癌密切关系的发现使得越来越多的英国男性不吸烟，现代肺癌手术的技术进步使得肺癌存活率大大增加，故此男性肺癌死亡率的下降是健康文明的社会进步，不仅英国如此，中国等发展中国家近年来也体现出这种发展特点。1971 年每十万人中死亡男性 107 人，女性 18 人，到 2009 年每十万人中死亡男性只有 49 人，而女性却达到了 31 人。38 年时间，男性死亡率下降了一多半，而女性死亡率上升了 50%。造成这种现象的原因是多方面的，一方面是人口统计学因素，即老龄化社会的到来，女性老人增加导致肺癌死亡率高，其次是大量的新移民女性到英国后由于工作压力开始吸烟，故而在整体上导致英国女性肺癌死亡率的小幅度上升。

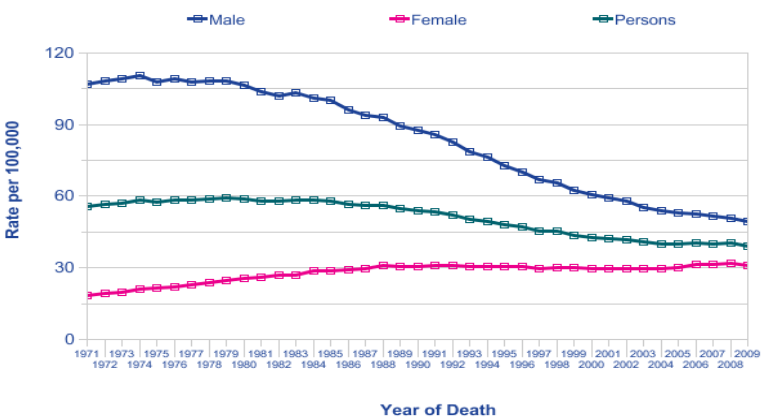


图 2-18　英国患者死亡率曲线

2.4.6　老龄肺癌死亡率数据的三维分析

如图 2-19 所示，就年龄而言，年纪越大，手术后死亡率越高。这是因为老年人器官功能退化，

术后并发症多发的原因，其中 70～90 岁高龄组最为明显。就手术方式而言，肺全切、肺段切、肺叶切、肺边切的死亡率也是不同的，10 万例肺癌手术中，Pneumonectomy 死亡率最高为 7.2%，Segmental/wedge 最低为 1.5%，Bilobectomy 的 4.6%高于 Lobectomy 的 2.5%，这深刻地表明肺癌生长部位，手术切除部位与方法对患者存活率有重大影响，老人肺全切可能导致呼吸系统衰竭。

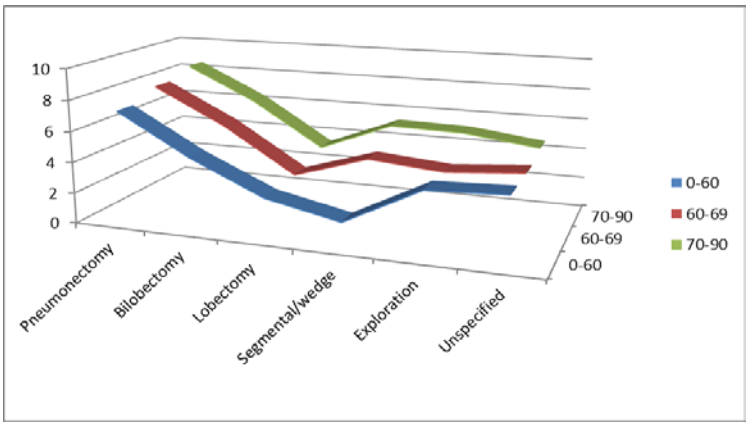


图 2-19　10 万例肺癌手术方式、术后死亡率、年龄的关系

如图 2-20、图 2-21 所示，在老年肺癌死亡率研究过程中，我们发现影响死亡率的主要因素是年龄、临床分期、病灶位置、手术方式及地区特点，本研究模型中暂不考虑并发症因素。

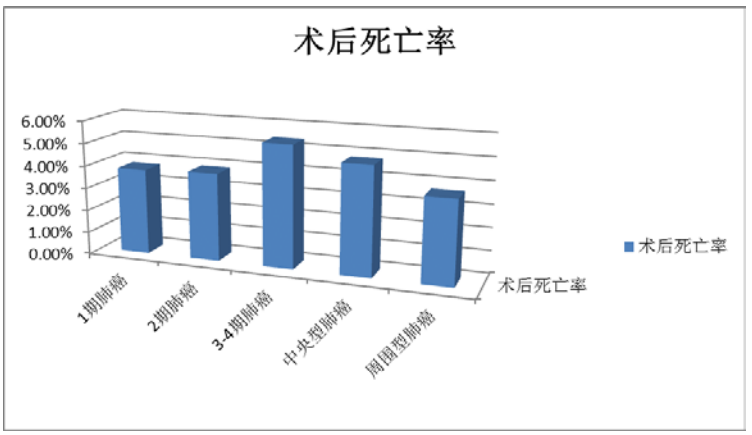


图 2-20　1 0 万例肺癌手术后死亡率 POM% Postoperative mortality (POM)

资料来源：邵学杰. 老年肺癌的最新统计学研究.

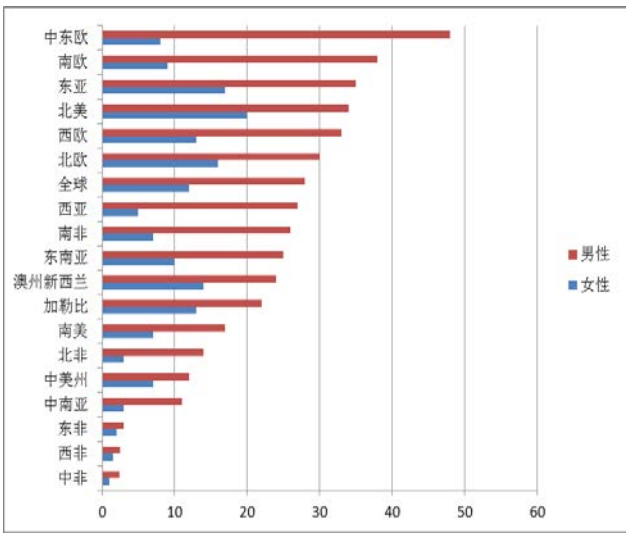


图 2-21　全球肺癌死亡率统计（每 100,000 人）

资料来源：WTO 2010

影响术后死亡率的因素首先是年龄，70～90 岁老龄患者术后死亡率最高，再一次证明老年肺癌的最佳手术时间不能超过 70 岁。其次手术的方式对生存也有重大影响，肺全切手术死亡率最高，肺段切次之，肺叶切死亡率最低，其中医学界经统计发现右肺全切是肺癌手术中风险比较大，术后死亡率达到 18%的高风险手术，特别是老年人的右肺全切产生的合并症最多。临床分期中，1 期、2 期死亡率差别不大，3-4 期高一些，10 万例手术中达到了 5.4%，就位置而言，中央型肺癌比周围型肺癌术后死亡率高。

值得我们关注的是全球地理分布因素对肺癌死亡率的影响。非洲国家肺癌死亡率很低的现象值得我们进一步研究，经济发展水平与碳排放标准、空气环境质量息息相关。已经揭示的规律显示，肺癌与吸烟、环境因素、职业病污染有很大的关系，非洲国家还没有完成工业化，相对低的碳排放环境，吸烟减少等都有可能是肺癌发病率低、死亡率低的重要因素。相对而言，中国所处的东亚地区肺癌死亡率居世界第三位，中国是肺癌高发国家。中国近 30 年的工业化加速过程中吸烟、空气污染、人们对肺癌的科学认识不足、早期发现难等都造成每十万人口死亡的高死亡率的现象。肺癌，特别是老年肺癌的一个重要特点就是发现难，很多老人发现肺癌时已经是晚期，肺癌不易发现的特性期待早期主动体检的行为倡导。

结论：通过对老年肺癌的网上数据分析，得出以下结论。

① 80 岁以上老年肺癌的手术禁忌年龄不是问题，只要心肺功能好，都可以承受心胸外科手术。

② 老年肺癌的预后与围术期及康复期的家庭家人护理密不可分。

③ 老年肺癌的手术方法对预后有重要的影响，Pneumonectomy 死亡率最高为 7.2%，Segmental/wedge 最低为 1.5%，Bilobectomy 的 4.6%高于 Lobectomy 的 2.5%，手术部位对五年生存

期有重要的影响。

④ 数据挖掘表明，开放式手术与胸腔镜在老年肺癌手术的预后上并没有太大的差别，小切口的开放式手术仍然受到外科医生的青睐。

⑤ 老年肺癌的生长部位也对五年预后生存期有重要的影响。

⑥ 对老年人而言，右肺全切是一个高风险死亡率的事件。

⑦ 由于老年肺癌的发现是一个难题，约有一半以上的肺癌患者发现了也已经是中晚期，早期筛查仍然是最好的方法。

⑧ 数据分析发现，由于女性外来移民的影响，欧洲的女性吸烟者的增加导致女性肺癌患者增加。

⑨ 数据挖掘表明，非洲的肺癌死亡率最低说明环境因素与肺癌息息相关。

老年肺癌的研究表明，数据技术也可以总结临床经验，缩短医生的专业学习曲线。数据挖掘甚至于可以对手术的手法、手术部位、手术方案这些传统观念上很难量化的属性提出循证医学的有力证据。

这就是医学数据挖掘的力量。

2.5 临床医学与数据挖掘的边缘学科

2.5.1 几个实例

例 1：某地用 A、B 和 C 三种方案治疗血红蛋白含量不满 10g 的婴幼儿贫血患者，A 方案为每公斤体重每天口服 2.5%硫酸亚铁 1ml，B 方案为每公斤体重每天口服 2.5%硫酸亚铁 0.5ml，C 方案为每公斤体重每天口服 3g 鸡肝粉，治疗一月后，记录下每名受试者血红蛋白的上升克数，资料见表 2-16，问三种治疗方案对婴幼儿贫血的疗效是否相同?

表 2-16　A、B、C 三种方案治疗婴幼儿贫血的疗效观察

治疗方案	血红蛋白增加量（g）									
A	1.8	0.5	2.3	3.7	2.4	2.0	1.5	2.7	1.1	0.9
（n=20）	1.4	1.2	2.3	0.7	0.5	1.4	1.7	3.0	3.2	2.5
B	0.2	0.5	0.3	1.9	1.0	2.4	–0.4	2.0	1.6	2.0
（n=19）	0.0	1.6	3.0	1.6	0.0	3.0	0.7	1.2	0.7	
C	2.1	1.9	1.7	0.2	2.0	1.5	0.9	1.1	–0.2	1.3
（n=20）	–0.7	1.3	1.1	0.2	0.7	0.9	0.8	–0.3	0.7	1.4

方差分析计算表如表 2-17 所示。

表 2-17　完全随机设计方差分析计算表

变异来源	*SS*	*v*	*MS*	*F*
总变异	$\sum_{i=1}^{k}\sum_{j=1}^{x_i}(x-\bar{x})^2=\sum x^2-C$	*N*-1		
组间（处理）	$\sum_{i=1}^{k}n_i(\bar{X}_i-\bar{X})^2=\sum_{i=1}^{k}\frac{(\sum X_1)^2}{n_1}-C$	*k*-1	$SS_{TR}v_{TR}$	MS_{TR}/MS_e
组内（误差）	SS_T-SS_{TR}	*N*-*k*	SS_e/v_e	

$$C=(\sum X)^2/N$$

F 检验步骤：

（1）建立假设

H_0：$m_A=m_B=m_C$，三种治疗方案治疗婴幼儿贫血的疗效相同；

H_1：三种治疗方案治疗婴幼儿贫血的疗效不全相同或全不相同。

（2）确立检验水准

a=0.05。

（3）计算检验统计量

① 计算各组基础数据：$\sum X_1$ 和 $\sum X_1^2$ 以及总的 $\sum X$ 和 $\sum X^2$。如表 2-18 所示。

表 2-18　方差总和表

	A	B	C	总和
$\sum X_1$	36.80	23.30	18.60	78.70
$\sum X_1^2$	83.56	47.01	28.86	159.43
n_i	20	19	20	59

② 分别计算 SS_T、SS_{TR}、和 SS_e。

$C=(78.70)^2/59=104.9778$

总变异：$SS_T=159.43-104.9778$

$=54.4522$

组间变异：$SS_{TR}=\frac{36.80^2}{20}+\frac{23.30^2}{19}+\frac{18.60^2}{20}-104.9778$

$=8.6054$

组内变异：$SS_e=54.4522-8.6054$

$=45.8468$

③ 列出方差分析表，如表 2-19 所示。

表 2-19　方差分析表

变异来源	*SS*	*n*	*MS*	*F*
总	54.4522	58		

续表

组间	8.6054	2	4.3027	5.2555
组内（误差）	45.8468	56	0.8187	

（4）确定 P 值

该 F 值分子的自由度 $n_{TR}=2$，分母的自由度 $n_e=56$，查 F 界值表得 $F_{0.05(2,56)}=3.16$，$F>F_{0.05(2,56)}$，则 $P<0.05$。

（5）结论

按 $a=0.05$ 水准，拒绝 H_0，接受 H_1，差别有统计学意义。故可认为三种治疗方案的治疗效果不一样。

例 2：为探讨一氧化氮（NO）在肾缺血再灌流过程中的作用，将 36 只雄性 SD 大鼠随机等分为 3 组给予不同处理后，测得 NO 数据见表 2-20，问各组 NO 水平是否相同?

表 2-20　三组大鼠肾组织液中 NO 水平（μmol·L^{-1}）

	正常对照组	肾缺血 60min 组	肾缺血 60min　再灌流组	合　　计
	437.98	322.75	284.04	
	285.75	464.51	194.90	
	369.93	322.34	197.53	
	344.53	282.52	227.57	
	378.96	278.47	184.42	
	300.92	348.47	223.17	
	271.70	354.10	363.43	
	417.97	302.21	390.38	
	287.10	269.65	332.68	
	363.51	322.98	355.99	
	309.60	288.76	219.72	
	338.83	386.67	143.17	
n_i	12	12	12	36
$\sum X$	4106.78	3943.43	3117.00	11167.21
$\sum X^2$	1436935.8666	1329275.5339	883943.8218	3650155.2223

（1）建立检验假设

H_0：$\mu_1=\mu_2=\mu_3$，三组大鼠 NO 含量总体均值相等；

H_1：三组大鼠 NO 含量总体均值不全相等或全不相等。

（2）确立检验水准

$\alpha=0.05$。

（3）计算检验统计量

$C=(\sum X)^2/N$ =（11167.21）^{2}/36

=3464071.6440

总变异：$SS_T=\sum X^2-C$

=186083.5783

组间变异：$SS_{TR}=\sum_{i=1}^{k}\frac{(\sum X_i)^2}{ni}-C$

=46925.9504

列出方差分析表，如表 2-21 所示：

表 2-21　方差分析表

变异来源	*SS*	*v*	*MS*	*F*
总	186083.578	35		
组间	46925.950	2	23462.975	5.564
组内（误差）	139157.628	33	4216.898	

（4）确定 *P* 值

该 *F* 值分子的自由度ν_{TR}=2，分母的自由度ν_e=33，查 *F* 界值表（附表 4）得 $F_{0.05(2,33)}$=3.30，$F>F_{0.05(2,33)}$，则 $P<0.05$。

（5）结论

按α=0.05 水准，拒绝 H_0，接受 H_1，差别有统计学意义。故可认为三组大鼠的 NO 水平不同。

例 3：在抗癌药筛选试验中，拟用 20 只小白鼠按不同窝别分为 5 组，分别观察三种药物对小白鼠肉瘤（S180）的抑瘤效果，资料见表 2-22，问三种药物有无抑瘤作用?

表 2-22　三种药物抑瘤效果的比较（瘤重：g）

窝别（配伍组）	对照	A	B	C	配伍组合计($\sum X_i$)
Ⅰ	0.80	0.36	0.17	0.28	1.61
Ⅱ	0.74	0.50	0.42	0.36	2.02
Ⅲ	0.31	0.20	0.38	0.25	1.14
Ⅳ	0.48	0.18	0.44	0.22	1.32
Ⅴ	0.76	0.26	0.28	0.13	1.43
处理组合计 $\sum X_i$	3.09	1.50	1.69	1.24	7.52($\sum X$)
$\sum X_i^2$	2.0917	0.5196	0.6217	0.3358	3.5688($\sum X^2$)

解题思路：

本例的主要目的是研究三种药物对小白鼠肉瘤（S180）的抑瘤效果，药物是处理因素。但是，不同窝别的小白鼠对肉瘤生长的反应若有差别，这种差别必定影响对药物效应的分析，因此在实验设计时可将不同窝别的小白鼠视为干扰因素，并作为区组，则在数据分析时就可以将处理因素的作

用与干扰因素的影响区分开，提高检验功效。

两因素方差分析的原理类似于单因素方差分析，前者仅在后者的基础上，从误差中再分离出区组效应，使误差减少，达到提高检验功效的目的。方差分析计算表如表 2-23 所示

表 2-23　随机区组设计方差分析计算表

变异来源	*SS*	*v*	*MS*	*F*
总变异	$\sum X^2 - C$	*N*-1 或 *kb*-1		
处理（A）	$\sum_{i=1}^{k} \frac{(\sum X_i)^2}{b} - C$	*k*-1	SS_A/v_A	MS_A/MS_e
区组（B）	$\sum_{j=1}^{k} \frac{(\sum X_i)^2}{k} - C$	*b*-1	SS_B/v_B	MS_B/MS_e
误差（e）	SS_T-SS_A-SS_B	*N*-*k*-*b* 或 (*k*-1)(*b*-1)	SS_e/v_e	

$C = (\sum x)^2 / N$

k 为因素 A 的水平数，*b* 为因素 B 的水平数。随机区组设计因素 A 每个水平的观察例数恰好等于因素 B 的水平数 *b*，而因素 B 每个水平的观察例数恰好等于因素 A 的水平数 *k*。

F 检验步骤：

（1）建立检验假设及设立检验水准

实验因素：

H_0：三种药物对小白鼠肉瘤（S180）的抑瘤效果与对照组相同，即$\mu_{对照}=\mu_A=\mu_B=\mu_C$；

H_1：三种药物对小白鼠肉瘤（S180）的抑瘤效果与对照组不全同或全不同。

α =0.05。

干扰因素：

H_0：5 个窝别小白鼠对肉瘤生长的反应相同；

H_1：5 个窝别小白鼠对肉瘤生长的反应不全相同或全不相同。

α=0.05。

（2）计算检验统计量

① 计算 *C* 值：

$$C = \frac{(\sum x)^2}{kb} = \frac{(7.52)^2}{5 \times 4} = 2.82752$$

② 计算总的离均差平方和 SS_T：

$$SS_X = \sum X^2 - C = 3.5688 - 2.82752 = 0.74128$$

③ 计算处理组间离均差平方和 SS_A：

$$SS_A=\sum\frac{(\sum X_i)^2}{k}-C=\frac{(3.09)^2}{5}+\frac{(1.50)^2}{5}+\frac{(1.69)^2}{5}+\frac{(1.24)^2}{5}-2.82752=0.41084$$

④ 计算区组间离差平方和 SS_B：

$$SS=\sum\frac{(\sum X_i)^2}{k}-C=\frac{(1.61)^2}{4}+\frac{(2.02)^2}{4}+\frac{(1.14)^2}{4}+\frac{(1.32)^2}{4}+\frac{(1.43)^2}{4}-2.82752=0.11233$$

⑤ 计算误差离均差平方和 SS_e：

$$SS_e=SS_T-SS_A-SS_B=0.74128-0.41084-0.11233=0.21811$$

⑥ 计算自由度：

总自由度　　$\nu_T=N-1=20-1=19$

处理组自由度　$\nu_A=k-1=4-1=3$

区组自由度　$\nu_B=b-1=5-1=4$

误差自由度　　$\nu_e=\nu_T-\nu_A-\nu_B=19-3-4=12$

⑦ 列出方差分析表，如表 2-24 所示。

表 2-24　两因素方差分析表

变异来源	*SS*	ν	*MS*	*F*	*P*
总	0.74128	19			
处理	0.41084	3	0.13695	7.53	< 0.05
区组	0.11233	4	0.02808	1.54	> 0.05
误差	0.21811	12	0.01818		

（3）确定 *P* 值

处理组按 $\nu_1=3$、$\nu_2=12$ 查 *F* 界值表得 $F_{0.05,(3,12)}=3.49<F_A$，则 $P<0.05$；

区组按 $\nu_1=4$、$\nu_2=12$ 查 *F* 界值表得 $F_{0.05,(4,12)}=3.26>F_B$，则 $P>0.05$。

（4）结论

对于处理组间，按α =0.05 的水准拒绝 H_0，接受 H_1，差别有统计学意义。可认为三种药物对小白鼠肉瘤（S180）的抑瘤效果与对照组不同。

对于区组间，按α =0.05 的水准不拒绝 H_0，差别无统计学意义。即各窝小白鼠对肉瘤生长的反应相同。

例 4：为比较不同产地石棉毒性的大小，取体重 200g～220g 的雌性 Wistar 大鼠 36 只，将月龄相同、体重相近的三只分为一组。每组的 3 只大鼠随机分别接受不同产地的石棉处理后，以肺泡巨噬细胞（PAM）存活率（%）评价石棉毒性大小。结果见表 2-25。问不同产地石棉毒性是否相同?

表 2-25　经不同产地石棉处理后大鼠的巨噬细胞存活率（%）

区组号（因素 B）	石棉产地（因素 A）			合　计
	甲　地	乙　地	丙　地	
1	50.88	44.01	66.97	161.86
2	48.02	66.27	71.92	186.21
3	45.26	59.99	69.89	175.14
4	38.38	52.49	67.05	157.92
5	52.70	60.69	56.35	169.74
6	60.22	66.12	70.08	196.42
7	44.49	55.36	86.60	186.45
8	49.31	53.39	68.20	170.90
9	46.23	52.34	63.36	161.93
10	51.16	55.16	66.12	172.44
11	42.48	58.64	70.02	171.14
12	53.47	61.08	67.24	181.79
$\sum X$	582.60	685.54	823.80	2091.94
$\sum X^2$	28648.9112	39604.4626	57085.4728	125338.8466

（1）建立检验假设及设立检验水准

实验因素：

H_0：三种产地石棉导致 PAM 存活率总体均数相等，即$\mu_1=\mu_2=\mu_3$；

H_1：三种产地石棉导致 PAM 存活率总体均数不等或不全相等。

α=0.05。

干扰因素：

H_0：不同区组动物的 PAM 存活率总体均数相等；

H_1：不同区组动物的 PAM 存活率总体均数不等或不全相等。

α=0.05。

（2）计算检验统计量

① 计算 C 值：

$$C=\frac{(\sum X)^2}{N}=\frac{(2091.94)^2}{36}=121561.4712$$

② 计算总的离均差平方和 SS_T：

$$SS_X=\sum X^2-C=125338.8466-121561.4712=3777.3754$$

③ 计算处理组间离均差平方和 SS_A：

$$SS_A=\sum\frac{(\sum X_i)^2}{b}-C=\frac{(582.6)^2}{12}+\frac{(685.54)^2}{12}+\frac{(823.8)^2}{12}-C=2441.3864$$

④ 计算区组间离差平方和 SS_B：

$$SS_B=\sum\frac{(\sum X_i)^2}{k}-C=\frac{(161.86)^2}{3}+\frac{(186.21)^2}{3}+\cdots+\frac{(181.79)^2}{3}-C=485.8116$$

⑤ 计算误差离均差平方和 SS_e：

$$SS_e=SS_T-SS_A-SS_B=850.1774$$

⑥ 计算自由度：

总自由度　　$\nu_T=N-1=36-1=35$

处理组自由度　$\nu_A=k-1=3-1=2$

区组自由度　　$\nu_B=b-1=12-1=11$

误差自由度　　$\nu_e=\nu_T-\nu_A-\nu_B=35-2-11=22$

⑦ 列出方差分析表，如表 2-26 所示。

表 2-26　两因素方差分析表

变异来源	*SS*	*v*	*MS*	*F*	*P*
总变异	3777.375	35			
处理（石棉）	2441.386	2	1220.693	31.588	< 0.05
区组（动物）	485.812	11	44.165	1.143	> 0.05
误差	850.177	22	38.644		

（3）确定 P 值

处理组按 $\nu_1=2$、$\nu_2=22$ 查 F 界值表得 $F_{0.05,(2,22)}=3.44<F_A$，则 $P<0.05$；

区组按 $\nu_1=11$、$\nu_2=22$ 查 F 界值表得 $F_{0.05,(11,22)}=2.26>F_B$，则 $P>0.05$。

（4）结论

对于处理组间，按 $\alpha=0.05$ 的水准拒绝 H_0，接受 H_1，差别有统计学意义。可认为三种产地石棉导致 PAM 存活率总体均数不等。

对于区组间，按 $\alpha=0.05$ 的水准不拒绝 H_0，差别无统计学意义。即不能认为不同区组大鼠间的 PAM 存活率不同。

资料来源：顾坚. 医学统计学. 南通大学.

2.5.2　医学统计学与医学数据挖掘的区别

数据挖掘来源于统计分析，而又不同于统计分析。数据挖掘不是为了替代传统的统计分析技术。相反，数据挖掘是统计分析方法的扩展和延伸。很多情况下，数据挖掘的本质是很偶然地发现非预

期但很有价值的信息。这说明数据挖掘过程本质上是实验性的。这和确定性的分析是不同的（实际上，一个人是不能完全确定一个理论的，只能提供证据和不确定的证据）。确定性分析着眼于最适合的模型——建立一个推荐模型，这个模型也许不能很好地解释观测到的数据。很多、或许是大部分统计分析提出的是确定性的分析。然而，实验性的数据分析对于统计学并不是新生事物，或许这是统计学家应该考虑作为统计学的另一个基石，而这已经是数据挖掘的基石。所有这些都是正确的，但事实上，数据挖掘所遇到的数据集按统计标准来看都是巨大的。在这种情况下，统计工具可能会失效：百万个偶然因素可能就会使其失效。

统计学和数据挖掘有着共同的目标：发现数据中的结构。事实上，由于它们的目标相似，一些人（尤其是统计学家）认为数据挖掘是统计学的分支。这是一个不切合实际的看法。因为数据挖掘还应用了其他领域的思想、工具和方法，尤其是计算机学科，例如数据库技术和机器学习，而且它所关注的某些领域和统计学家所关注的有很大不同。

统计学和数据挖掘研究目标的重迭自然导致了迷惑。事实上，有时候还导致了反感。统计学有着正统的理论基础（尤其是经过本世纪的发展），而现在又出现了一个新的学科，有新的主人，而且声称要解决统计学家们以前认为是他们领域的问题。这必然会引起关注。更多的是因为这门新学科有着一个吸引人的名字，势必会引发大家的兴趣和好奇。把“数据挖掘”这个术语所潜在的承诺和“统计学”作比较的话，统计的最初含义是“陈述事实”，以及找出枯燥的大量数据背后的有意义的信息。当然，统计学的现代含义已经有很大不同的事实。而且，这门新学科同商业有特殊的关联（尽管它还有科学及其他方面的应用）。区分它们的异同，并关注与数据挖掘相关联的一些难题。首先，我们注意到“数据挖掘”对统计学家来说并不陌生。例如，Everitt 定义它为：“仅仅是考察大量的数据驱动的模型，从中发现最适合的”。统计学家因而会忽略对数据进行特别的分析，因为他们知道太细致的研究却难以发现明显的结构。尽管如此，事实上大量的数据可能包含不可预测的但很有价值的结构。而这恰恰引起了注意，也是当前数据挖掘的任务。由于统计学基础的建立时代在计算机的发明和发展之前，所以常用的统计学工具包含很多可以手工实现的方法。因此，对于很多统计学家来说，1000 个数据就已经是很大的了。但这个“大”对于英国大的信用卡公司每年 350 000 000 笔业务或 AT&T 每天 200 000 000 个长途呼叫来说相差太远了。很明显，面对这么多的数据，则需要设计不同于那些“原则上可以用手工实现”的方法。这意味着计算机（正是计算机使得大数据可能实现）对于数据的分析和处理是关键的。分析者直接处理数据将变得不可行。相反，计算机在分析者和数据之间起到了必要的过滤作用。这也是数据挖掘特别注重准则的另一原因。尽管有必要，把分析者和数据分离开很明显导致了一些关联任务。这里就有一个真正的危险：非预期的模式可能会误导分析者，这一点我下面会讨论。

我不认为在现代统计中计算机不是一个重要的工具。它们确实是，并不是因为数据的规模。对数据的精确分析方法如 Bootstrap 方法、随机测试、迭代估计方法以及比较适合的复杂的模型正是有了计算机才是可能的。计算机已经使得传统统计模型的视野大大地扩展了，还促进了新工具的飞速发展。

下面来关注一下歪曲数据的非预期模式出现的可能性。这和数据质量相关。所有数据分析的结

论依赖于数据质量。GIGO 的意思是垃圾进，垃圾出，它的引用到处可见。一个数据分析者，无论他多聪明，也不可能从垃圾中发现宝石。对于大的数据集，尤其是要发现精细的小型或偏离常规的模型的时候，这个问题尤其突出。当一个人在寻找百万分之一的模型的时候，第二个小数位的偏离就会起作用。一个经验丰富的人对于此类最常见的问题会比较警觉，但出错的可能性太大了。

此类问题可能在两个层次上产生。第一个是微观层次，即个人记录。例如，特殊的属性可能丢失或输入错误。我知道一个案例，由于挖掘者不知道，丢失的数据被记录为 99 而作为真实的数据处理。第二个是宏观层次，整个数据集被一些选择机制所歪曲。交通事故为此提供了一个好的示例。越严重的、致命的事故，其记录越精确，但小的或没有伤害的事故的记录却没有那么精确。事实上，很高比例的数据根本没有记录。这就造成了一个歪曲的映象，可能会导致错误的结论。

统计学很少会关注实时分析，然而数据挖掘问题常常需要这些。例如，银行事务每天都会发生，没有人能等三个月得到一个可能的欺诈的分析。类似的问题发生在总体随时间变化的情形。我的研究组有明确的例子显示银行债务的申请随时间、竞争环境、经济波动而变化。

至此，我们已经论述了数据分析的问题，说明了数据挖掘和统计学的差异，尽管有一定的重叠。但是，数据挖掘者也不可持完全非统计的观点。首先来看一个例子：获得数据的问题。统计学家往往把数据看成一个按变量交叉分类的平面表，存储于计算机等待分析。如果数据量较小，可以读到内存，但在许多数据挖掘问题中这是不可能的。更糟糕的是，大量的数据常常分布在不同的计算机上。或许极端的是，数据分布在全球互联网上。此类问题使得获得一个简单的样本不大可能（先不管分析“整个数据集”的可能性，如果数据是不断变化的这一概念可能是不存在的，例如电话呼叫）。

当描述数据挖掘技术的时候，我发现依据以建立模型还是模式发现为目的可以很方便地区分两类常见的工具。我已经提到了模型概念在统计学中的核心作用。在建立模型的时候，尽量要概括所有的数据，以及识别、描述分布的形状。这样的“全”模型的例子包括对一系列数据的聚类分析、回归预测模型，以及基于树的分类法则等。相反，在模式发现中，则是尽量识别小的（但不一定不重要）偏差，发现行为的异常模式。例如 EEG 轨迹中的零星波形、信用卡使用中的异常消费模式，以及不同于其他特征的对象。很多时候，这第二种实验是数据挖掘的本质——试图发现渣滓中的金块。然而，第一类实验也是重要的。当关注的是全局模型的建立的话，样本是可取的（可以基于一个十万大小的样本发现重要的特性，这和基于一个千万大小的样本是等效的，尽管这部分地取决于我们想法的模型的特征）。然而，模式发现不同于此。仅选择一个样本的话可能会忽略所希望检测的情形。

尽管统计学主要关注的是分析定量数据，数据挖掘的多来源意味着还需要处理其他形式的数据。特别的，逻辑数据越来越多——例如当要发现的模式由连接的和分离的要素组成的时候。类似的，有时候会碰到高度有序的结构。分析的要素可能是图像、文本、语言信号，或者甚至完全是（例如，在交替分析中）科学研究资料。

数据挖掘有时候是一次性的实验。这是一个误解。它更应该被看作是一个不断的过程（尽管数据集有时是确定的）。从一个角度检查数据可以解释结果，以相关的观点检查可能会更接近等等。关键是，除了极少的情形下，很少知道哪一类模式是有意义的。数据挖掘的本质是发现非预期的模

式——同样非预期的模式要以非预期的方法来发现。

与把数据挖掘作为一个过程的观点相关联的是认识到结果的新颖性。许多数据挖掘的结果是我们所期望的——可以回顾。然而，可以解释这个事实并不能否定挖掘出它们的价值。没有这些实验，可能根本不会想到这些。实际上，只有那些可以依据过去经验形成合理的解释的结构才会是有价值的。

显然在数据挖掘存在着一个潜在的机会。在大数据集中发现模式的可能性当然存在，大数据集的数量与日俱增。然而，也不应就此掩盖危险。所有真正的数据集（即使那些是以完全自动方式搜集的数据）都有产生错误的可能。关于人的数据集（例如事务和行为数据）尤其有这种可能。这很好地解释了绝大部分在数据中发现的“非预期的结构”本质上是无意义的，而是因为偏离了理想的过程（当然，这样的结构可能会是有意义的：如果数据有问题，可能会干扰搜集数据的目的，最好还是了解它们）。与此相关联的是如何确保（和至少为事实提供支持）任何所观察到的模式是“真实的”，它们反映了一些潜在的结构和关联而不仅仅是一个特殊的数据集，由于一个随机的样本碰巧发生。在这里，记分方法可能是相关的，但需要更多的统计学家和数据挖掘工作者的研究。

原文出自 *Statistics and Data Mining：Intersecting Disciplines* 作者：David J. Hand

2.5.3 有关数据挖掘是边缘学科的几个实例

实例 1:

公元 1814 年，法国数学家普斯（1794—1827）在他的新作《概率的哲学讨论》一书中，讲述了一个有趣的统计。他根据全法国的统计资料，得出了几乎完全一致的男婴和女婴出生数的比值是 22：21，即在全体出生婴儿中，男婴 51.2，女婴 48.8。可悲的是，当他统计 1745 年～1784 年整整四十年间巴黎男婴出生率时却得到了另一个男女出生比例 25：24，男婴 51.02，与之前相比相差了 0.14。对于这千分之一点四的微小差异，普斯感到不解，他深信自然规律，他知道这千分之一点四的后面，一定有深层的原因。于是，他深入进行调查研究，发现：当时巴黎人“重女轻男”，有男婴被遗弃的现象。搞清楚了出生率的真相，经过复核计算，巴黎的男女婴的出生比率依然是 22：21。

这个故事充分说明了统计学与数据挖掘的区别，更阐述了数据挖掘的边缘性学科特点。首先是对全国出生人口的数据统计与记录，在大数据条件下 TB 级、PB 级的数据量依靠传统的手工统计或简单的计算机程序已经完全无法解决，这体现了数据挖掘需要计算机数据库，特别是大型分布式数据库的存储与处理技术，此外，海量数据的在线查询、机器学习等方法的运用说明了数据挖掘的学科交叉与边缘性特点。

实例 2:

德·梅勒是一位军人、语言学家、古典学者，同时也是一个有能力、有经验的赌徒，他经常玩骰子和纸牌。虽然他不是一个全职的数学家，但他经常从数学的角度提出和思考赌博中出现的一些有深度的问题，“点问题” 就是其中之一。这一次，德·梅勒的问题的形式是：假设两个赌博者（德·梅勒和他的一个朋友）每人出 30 个金币，两人各自选取一个点数，谁选择的点数首先被掷出 3 次，谁

就赢得全部的赌注。在游戏进行了一会儿后，德·梅勒选择的点数“5”出现了 2 次，而他的朋友选择的点数“3”只出现了一次。这时候，德·梅勒由于一个紧急事情必须离开，游戏不得不停止。他们该如何分配赌桌上的 60 个金币的赌注呢？德·梅勒的朋友认为，既然掷出他选择的点数的机会是德·梅勒的一半，那么他该拿到德·梅勒所得的一半，即他拿 20 个金币，德·梅勒拿 40 个金币。然而德·梅勒争执到：再掷一次骰子，对他来说最糟糕的事是他将失去他的优势，游戏是平局，每人都得到相等的 30 个金币；但如果掷出的是“5”，他就赢了，并可拿走全部的 60 个金币。在下一次掷骰子之前，他实际上已经拥有了 30 个金币，他还有 50%的机会赢得另外 30 个金币，所以，他应分得 45 个金币。

他们对这一问题的看法和计算方法不一致，为此而争论不休。后来德·梅勒把这个问题告诉了帕斯卡，帕斯卡对此也很感兴趣，又写信告诉了费马。于是在这两位伟大的法国数学家之间开始了具有划时代意义的通信。在通信中，两人用不同的方法正确地解决了这个问题。在 1654 年 7 月 29 日帕斯卡写给费马的信中，他提到了这个问题和可能的解决方法，“你的解法非常正确，是给我印象最深的一个，但这些组合太过麻烦。我发现了另一种更为简洁的实在可行的解法。”在 1654 年 10 月 21 日他写给费马的信中提到，当他们互不赞同的时候，能这样通信，保持一致是鼓舞人心的。他说：“先生，您的最后一封信让我非常满意，您有关‘点问题’的解法我很钦佩。更是因为我非常理解它完全是属于你的，它与我的解法完全不同，然而却轻易地得到了同样的结果，现在我们又开始和睦了。”在 1654 年 7 月和 10 月的通信中，他们还联系“点问题”思考了其他的问题，比如当两人的技艺不等时，或超过 2 人参加游戏的赌金的分配问题。尤其是帕斯卡的研究更有效地推动了数学概率理论的发展，他的组合方法具有一般性。他的工作中还蕴涵了概率论中另一重要的思想——数学期望的思想。在十七世纪弥漫着浓重的宗教神学气质的精神环境中，身为神学家的帕斯卡也结合了宗教和数学两种思想于概率的思考中。帕斯卡在他的著作《思想录》中曾经提出一个以“帕斯卡赌注”闻名的问题：“我们既不知道上帝的存在，也不知道上帝的本质。然而我们将倾向于哪一边呢……这里进行的是一场赌博…… 让我们来权衡一下在上帝存在的赌注中的得失。让我们估计这两种可能性，如果你赢了，你赢得所有；如果你输了，你却一无所失。因此，你就不必迟疑去赌上帝的存在吧。”这个论述中已包含了比较明确的数学期望的思想，这种思想成为以后惠更斯（Christian Huggens）和维特（De. Witt）的概率论工作中的一个基本思想，并在以后相当长的时间里在古典概率论的研究中起着重要的作用。

帕斯卡和费马正确解决了“点问题”的这一事件被伊夫斯（Howard Eves）称为“数学史上的一个里程碑”。在概率论的历史上，一般的传统观点则把这一事件看作数学概率论的起始标志。之所以不把卡尔达诺的著作作为概率论的起源的始点，有这样几个原因：在卡尔达诺的著作中只有一小部分内容是处理机会（chance）的计算的。就像卡尔达诺的大多数作品一样，这种处理似乎只是零碎的和模糊的，混杂于卡尔达诺的个人的一些奇闻轶事、哲学思考、大量流行的赌博者常用的欺骗策略和精明的心理应用等建议之中，并且他的这本著作中所阐述的数学思想对数学家和一般的赌徒几乎都没有什么影响。因为对于当时的数学家而言，概率太游戏化了，而对赌徒而言，概率又太数学化了。而帕斯卡和费马的通信除了正确解决了一些问题和概念之外，还创造了一种研究的传统——

用数学方法（主要是组合数学的方法）研究和思考机会性游戏。这种传统统治这个领域达半个多世纪的时间。所以，综合考虑所有这些因素，这个事件赢得它在数学概率论历史中的标志性地位是当之无愧的。

这是一个流传甚广的经典的概率论大师的故事，其中或隐或现地总结了数理统计的方法论特点：抽样理论、参数设计、假设检验、方差分析、回归预测、推理演绎、归纳总结。对随机事件的假设检验是统计学的巨大进步，随着研究随机现象规律性的科学——概率论的发展，应用概率论的结果更深入地分析研究统计资料，通过对某些现象的频率的观察来发现该现象的内在规律性，并作出一定精确程度的判断和预测。将这些研究的某些结果加以归纳整理，逐步形成一定的数学模型，这些组成了数理统计的内容。

2.5.4 一个医学数据挖掘的案例

1. 数据挖掘一般步骤

① 分析问题：源数据数据库必须经过评估确认其是否符合数据挖掘标准，以决定预期结果，也就选择了这项工作的最优算法。

② 提取、清洗和校验数据：提取的数据放在一个结构上与数据模型兼容的数据库中。以统一的格式清洗那些不一致、不兼容的数据。一旦提取和清理数据后，浏览所创建的模型，以确保所有的数据都已经存在并且完整。

③ 创建和调试模型：将算法应用于模型后产生一个结构。浏览所产生结构中的数据，确认它对于源数据中“事实”的准确代表性，这是很重要的一点。虽然可能无法对每一个细节做到这一点，但是通过查看生成的模型，就可能发现重要的特征。

④ 查询数据挖掘模型的数据：一旦建立模型，该数据就可用于决策支持了。

⑤ 维护数据挖掘模型：数据模型建立好后，初始数据的特征，如有效性，可能发生改变。一些信息的改变会对精度产生很大的影响，因为它的变化影响作为基础的原始模型的性质。因而，维护数据挖掘模型是非常重要的环节。

聚类分析是数据挖掘采用的核心技术，成为该研究领域中一个非常活跃的研究课题。聚类分析基于“物以类聚”的朴素思想，根据事物的特征，对其进行聚类或分类。作为数据挖掘的一个重要研究方向，聚类分析越来越得到人们的关注。聚类的输入是一组没有类别标注的数据，事先可以知道这些数据聚成几簇也可以不知道聚成几簇。通过分析这些数据，根据一定的聚类准则，合理划分记录集合，从而使相似的记录被划分到同一个簇中，不相似的数据划分到不同的簇中。

2.特征选择与聚类分析算法

Relief 为一系列算法，它包括最早提出的 Relief 以及后来拓展的 ReliefF 和 RReliefF 算法，其中 RReliefF 算法是针对目标属性为连续值的回归问题提出的，下面仅介绍一下针对分类问题的 Relief 和 ReliefF 算法。

（1）Relief 算法

Relief 算法最早由 Kira 提出，最初局限于两类数据的分类问题。Relief 算法是一种特征权重算法（Feature weighting algorithms），根据各个特征和类别的相关性赋予特征不同的权重，权重小于某个阈值的特征将被移除。Relief 算法中特征和类别的相关性是基于特征对近距离样本的区分能力。算法从训练集 D 中随机选择一个样本 R，然后从和 R 同类的样本中寻找最近邻样本 H，称为 Near Hit，从和 R 不同类的样本中寻找最近邻样本 M，称为 NearMiss，然后根据以下规则更新每个特征的权重：如果 R 和 Near Hit 在某个特征上的距离小于 R 和 Near Miss 上的距离，则说明该特征对区分同类和不同类的最近邻是有益的，则增加该特征的权重；反之，如果 R 和 Near Hit 在某个特征的距离大于 R 和 Near Miss 上的距离，说明该特征对区分同类和不同类的最近邻起负面作用，则降低该特征的权重。以上过程重复 m 次，最后得到各特征的平均权重。特征的权重越大，表示该特征的分类能力越强，反之，表示该特征分类能力越弱。Relief 算法的运行时间随着样本的抽样次数 m 和原始特征个数 N 的增加线性增加，因而运行效率非常高。具体算法如下所示：

设训练数据集 D，样本抽样次数 m，特征权重的阀值 δ，输出是各个特征的权重 T：

① 置 0 所有特征权重，T 为空集；

② *for* i=1 *to* m *do*

- 随机选择一个样本 R；
- 从同类样本集中找到 R 的最近邻样本 H，从不同类样本集中找到最近邻样本 M；
- *for* A=1 *to* N *do*

$W(A)= W(A)–diff(A,R,H)/m+f\ diff(A,R.M)\ /m$

③ *for* A=1 *to* N *do*

if $W(A)\geqslant \delta$

把第 A 个特征加到 T 中

end

（2）ReliefF 算法

由于 Relief 算法比较简单，但运行效率高，并且结果也比较令人满意，因此得到广泛应用，但是其局限性在于只能处理两类别数据，因此 1994 年 Kononeill 对其进行了扩展，得到了 ReliefF 作算法，可以处理多类别问题。该算法用于处理目标属性为连续值的回归问题。ReliefF 算法在处理多类别问题时，每次从训练样本集中随机取出一个样本 R，然后从和 R 同类的样本集中找出 R 的 k 个近邻样本（near Hits），从每个 R 的不同类的样本集中均找出 k 个近邻样本（near Misses），然后更新每个特征的权重。如下式所示：

$$W(A)=W(A)-\sum_{j=1}^{k}diff(A,R,H_j)\ /mk)+$$

$$\sum_{Ceclass(R)}\left[\frac{P(C)}{1-p(Class(R))}\sum_{j=1}^{k}diff(A,R,M_j(C))\right]/mk)$$

在上式中，$diff(A,R_1,R_2)$表示样本 R_1 和样本 R_2 在特征 A 上的差，其计算公式，$M_j(C)$表示类 C 中的第 j 个最近邻样本。如下式所示：

$$diff(A,R_1,R_2)=\begin{cases}\dfrac{|R_1[A]-R_2[A]|}{\max(A)-\min(A)} & if\ A\ is\ continuous\\ 0 & if\ A\ is\ discrete\ and\ R_1[A]=R_2[A]\\ 1 & if\ A\ is\ discrete\ and\ R_1[A]\neq R_2[A]\end{cases}$$

ReliefF 算法具体的伪代码如下所示：

设训练数据集 D，样本抽样次数 m，特征权重的阀值 δ，最近邻样本个数 k；输出是各个特征的权重 T。

① 置所有特征权重为 0，T 为空集；

② for i=1 to m do

- 从 D 中随机选择一个样本 R；
- 从 R 的同类样本集中找到 R 的 k 个最近邻样本 H_j（$j=1,2,\cdots,k$），从每一个不同类样本集中找出 k 个最近邻 $M_j(C)$；

③ *for A*=1 *to N All feature do*

$$W(A)=W(A)-\sum_{j=1}^{k}diff(A,R,H_j)/(mk)+$$
$$\sum_{Ceclass(R)}\left[\frac{P(C)}{1-p(Class(R))}\sum_{j=1}^{k}diff(A,R,M_j(C))\right]/(mk)$$

end

Relief 系列算法运行效率高，对数据类型没有限制，属于一种特征权重算法，算法会赋予所有和类别相关性高的特征较高的权重，所以算法的局限性在于不能有效地去除冗余特征。

（3）K-Means 聚类算法

由于聚类算法是给予数据自然上的相似划法，要求得到的聚类是每个聚类内部数据尽可能地相似而聚类之间要尽可能地大差异。所以定义一种尺度来衡量相似度就显得非常重要了。一般来说，有两种定义相似度的方法。第一种方法是定义数据之间的距离，描述的是数据的差异。第二种方法是直接定义数据之间的相似度。下面是几种常见的定义距离的方法：

① Euclidean 距离，这是一种传统的距离概念，适合于 2、3 维空间。

② Minkowski 距离，是 Euclidean 距离的扩展，可以理解为 N 维空间的距离。

聚类算法有很多种，在需要时可以根据所涉及的数据类型、聚类的目的以及具体的应用要求来选择合适的聚类算法。下面介绍 K-Means 聚类算法：

K-Means 算法是一种常用的基于划分的聚类算法。K-Means 算法是以 k 为参数，把 n 个对象分

成 k 个簇，使簇内具有较高的相似度，而簇间的相似度较低。K-Means 的处理过程为：首先随机选择 k 个对象作为初始的 k 个簇的质心，然后将其余对象根据其与各个簇的质心的距离分配到最近的簇，最后重新计算各个簇的质心。不断重复此过程，直到目标函数最小为止。簇的质心由下列公式求得：

$$Z_j = \frac{1}{N_j} \sum_{xew_j} x$$

其中 N_j 表示属于 W_j 类的数据点的个数，属于某个簇的所有点的算术平均值即为该簇的质心。对象到质心的距离一般采用欧式距离。K-Means 算法的具体描述如下：

输入：聚类个数 k 以及包含 n 个数据对象的数据集；

输出：满足目标函数值最小的 k 个聚类。

算法流程：

① 从 n 个数据对象中任意选择 k 个对象作为初始聚类中心；

② 循环下述流程③到④，直到目标函数 E 取值不再变化；

③ 根据每个聚类对象的均值（中心对象），计算每个对象与这些中心对象的距离，且根据最小距离重新对相应对象进行划分；

④ 重新计算每个聚类的均值（中心对象）。

在具体实现时，为了防止步骤②中的条件不成立而出现无限循环，往往定义一个最大迭代次数。K-Means 尝试找出使平方误差函数值最小的 k 个划分。当数据分布较均匀，且簇与簇之间区别明显时，它的效果较好。面对大规模数据集，该算法是相对可扩展的，并且具有较高的效率。其中，n 为数据集中对象的数目，k 为期望得到的簇的数目，t 为迭代的次数。通常情况下，算法会终止于局部最优解。但也有例外，例如涉及有非数值属性的数据。其次，这种算法要求事先给出要生成的簇的数目 k，显然这对用户提出了过高的要求，并且由于算法的初始聚类中心是随机选择的，而不同的初始中心对聚类结果有很大的影响。另外，K-Means 算法不适用于发现非凸面形状的簇，或者大小差别很大的簇，而且它对于噪音和孤立点数据是敏感的。

3. 一个医学数据分析实例

（1）数据说明

本文实验数据来自著名的 UCI 机器学习数据库，该数据库有大量的人工智能数据挖掘数据，网址为 http://archive.ics.uci.edu/ml/。该数据库是不断更新的，也接受数据的捐赠。数据库种类涉及生活、工程、科学等各个领域，记录数也是从少到多，最多达几十万条。截止 2010 年底，数据库共有 199 个数据集，每个数据集合中有不同类型、时间的相关数据。可以根据实际情况进行选用。

本文选用的数据类型为：Breast Cancer Wisconsin（Original）Data Set，中文名称为：威斯康星州乳腺癌数据集。这些数据来源于美国威斯康星大学医院的临床病例报告，每条数据具有 11 个属性。下载下来的数据文件格式为“.data”，通过使用 Excel 和 Matlab 工具将其转换为 Matlab 默认的数据集保存，方便程序进行调用。

表 2-27 是该数据集的 11 个属性名称及说明。

表 2-27　威斯康星州乳腺癌数据集的 11 个属性名称及说明

属性名称	说　　明	特征编号
样品编号	病人身份证号码	无
块厚度	范围 1-10	1
细胞大小均匀性	范围 1-10	2
细胞形态均匀性	范围 1-10	3
边缘粘附力	范围 1-10	4
单上皮细胞尺寸	范围 1-10	5
裸核	范围 1-10	6
Bland 染色质	范围 1-10	7
正常核仁	范围 1-10	8
核分裂	范围 1-10	9
分类	分类属性：2 为良性 4 为恶性	10

对上述数据进行转换后，以及由数据说明可知，可以用于特征提取的有 9 个指标，样品编号和分类只是用于确定分类。本文的数据处理思路是先采用 ReliefF 特征提取算法计算各个属性的权重，剔除相关性最小的属性，然后采用 K-Means 聚类算法对剩下的属性进行聚类分析。

（2）数据预处理与程序

本文在转换数据后，首先进行了预处理，由于本文的数据范围都是 1～10，因此不需要归一化，但是数据样本中存在一些不完整数据，会影响实际的程序运行，经过程序处理，将这一部分数据删除。这些不完整的数据都是由于实际中一些原因没有登记或者遗失的，以“?”的形式表示。

本文采用 Matlab 软件进行编程计算。根据前文提到的 ReliefF 算法过程，先编写 ReliefF 函数程序，用来计算特征属性，再编写主程序，在主程序中调用该函数进行计算，并对结果进行分析、绘图，得到有用的结论。

程序统一在本书最后列出。

（3）乳腺癌数据集特征提取

本文采用第 2 小节中的 ReliefF 算法来计算各个特征的权重，权重小于某个阈值的特征将被移除，针对本文的实际情况，将对权重最小的 2～3 种剔除。由于算法在运行过程中，会选择随机样本 R，随机数的不同将导致结果权重有一定的出入，因此本文采取平均的方法，将主程序运行 20 次，然后将结果汇总求出每种权重的平均值。如下所示，列为属性编号，行为每一次的计算结果。

表 2-28 是特征提取算法计算的特征权重趋势数据，计算 20 次的结果趋势相同。

表 2-28　20 次计算的特征权重

	1	2	3	4	5	6	7	8	9
1	0.2207	0.1406	0.1434	0.1120	0.0644	0.2123	0.1163	0.1944	0.0375
2	0.2311	0.1488	0.1703	0.1470	0.0701	0.2491	0.1049	0.1724	0.0363
3	0.2111	0.1535	0.1568	0.1285	0.0755	0.2604	0.1243	0.2012	0.0693
4	0.2099	0.1865	0.1847	0.1694	0.0771	0.2337	0.1306	0.2219	0.0674
5	0.2436	0.1554	0.1689	0.1424	0.0628	0.2391	0.1309	0.2054	0.0479
6	0.2155	0.1460	0.1641	0.1220	0.0762	0.2366	0.1422	0.1936	0.0609
7	0.2436	0.1439	0.1759	0.1722	0.0752	0.2351	0.1351	0.2005	0.0431
8	0.2089	0.1443	0.1599	0.1571	0.0785	0.2399	0.1125	0.1759	0.0545
9	0.2273	0.1483	0.1615	0.1523	0.0674	0.2615	0.1399	0.2108	0.0394
10	0.2295	0.1314	0.1641	0.1439	0.0724	0.2517	0.1439	0.2068	0.0554
11	0.2120	0.1450	0.1240	0.1328	0.0703	0.2356	0.1234	0.1995	0.0535
12	0.2516	0.1385	0.1693	0.1484	0.0672	0.2580	0.1314	0.2062	0.0470
13	0.2507	0.1552	0.1642	0.1597	0.0785	0.2422	0.1224	0.1913	0.0347
14	0.2219	0.1615	0.1616	0.1293	0.0812	0.2361	0.1035	0.1870	0.0530
15	0.2075	0.1474	0.1490	0.1222	0.0738	0.2524	0.1299	0.1946	0.0319
16	0.2038	0.1462	0.1538	0.1510	0.0604	0.2200	0.1335	0.2172	0.0564
17	0.2302	0.1786	0.1707	0.1366	0.0757	0.2405	0.1280	0.2172	0.0679
18	0.2226	0.1097	0.1139	0.1205	0.0679	0.2401	0.1035	0.1616	0.0359
19	0.2083	0.1509	0.1701	0.1318	0.0870	0.2380	0.1210	0.2123	0.0467
20	0.2245	0.1559	0.1507	0.1373	0.0821	0.2330	0.1083	0.1884	0.0668

上述结果是否运行主程序所得的计算结果，看起来不直观，下面将其按照顺序绘图，可以直观显示各个属性权重的大小分布，如图 2-22 所示。特征属性权重均值如表 2-29 所示。

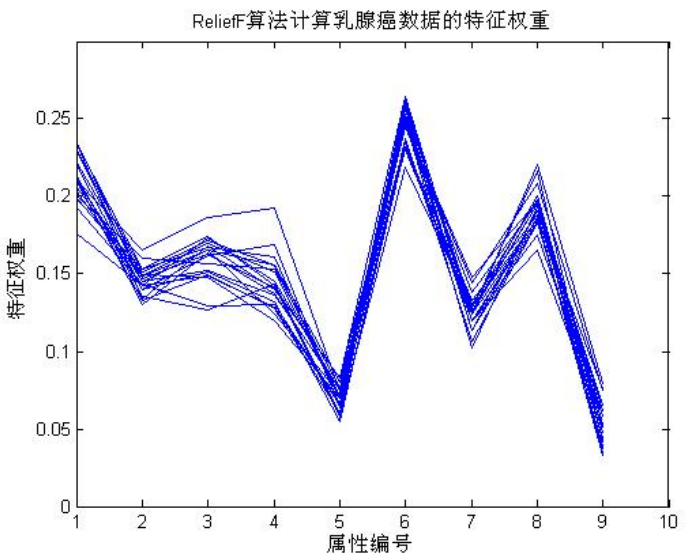


图 2-22　各个属性权重的大小分布

表 2-29　特征属性权重的均值

属性 1	0.2237	属性 2	0.1494
属性 3	0.1588	属性 4	0.1408
属性 5	0.0732	属性 6	0.2408
属性 7	0.1243	属性 8	0.1979
属性 9	0.0503		

按照从小到大顺序排列，可知各个属性的权重关系如下：

属性 9<属性 5<属性 7<属性 4<属性 2<属性 3<属性 8<属性 1<属性 6。

我们选定权重阈值为 0.02，则属性 9、属性 4 和属性 5 剔除。

从上面的特征权重可以看出，属性 6 裸核大小是最主要的影响因素，说明乳腺癌患者的症状最先表现在裸核大小上，将直接导致裸核大小的变化，其次是属性 1 和属性 8 等，后几个属性权重大小接近，但是从多次计算规律来看，还是能够说明其中不同的重要程度，下面是着重对几个重要的属性进行分析。图 2-23 是 20 次测试中，裸核大小（属性 6）的权重变化。

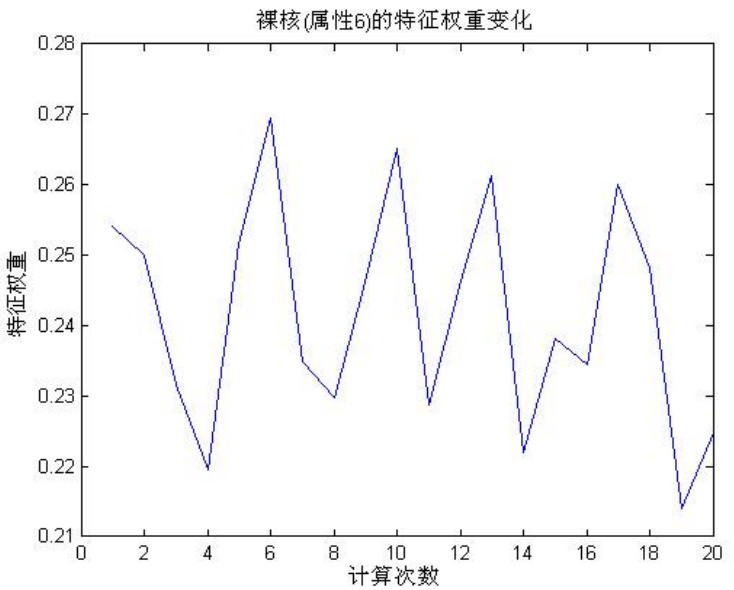


图 2-23　裸核（属性 6）的特征与权重变化

从图 2-23 中可以看到该属性权重大部分在 0.22～0.26 之间，是权重最大的一个属性。下面看看属性 1 的权重分布，如图 2-24 所示。

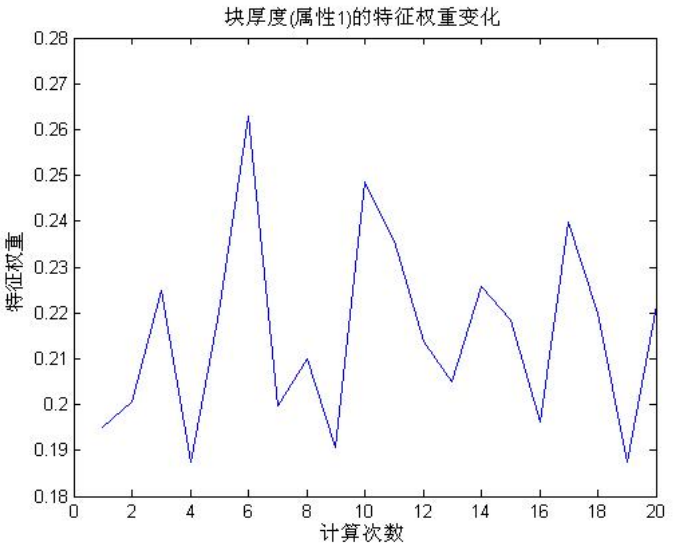


图 2-24　块厚度（属性 1）的特征与权重变化

块厚度属性的特征权重在 0.19～25 之间变动，也是权重极高的一个，说明该特征属性在乳腺癌患者检测指标中是相当重要的一个判断依据。进一步分析显示，在单独对属性 6 和属性 1 进行聚类分析，其成功率就可以达到 91.8%。本文将在下节中的 K-Means 算法中详细介绍。

（4）乳腺癌数据集聚类分析

上一节中通过 ReliefF 算法对数据集的分析，可以得到属性权重的重要程度，这些可以对临床诊断有一些参考价值，可以用来对实际案例进行分析，可以尽量地避免错误诊断，并提高诊断的速度和正确率。下面将通过 K-Means 聚类分析算法对数据进行分析。本小节将分为几个步骤来进行对比，确定聚类分析算法的结果以及与 ReliefF 算法结合的结果等。

① K-Means 算法单独分析数据集

下面将采用 K-Means 算法单独对数据集进行分析。Matlab 中已经包括了一些常规数据挖掘的算法，例如本文所用到的 K-Means 算法。该函数名为 kmeans，可以对数据集进行聚类分析。首先本文对乳腺癌数据集的所有属性列（除去身份信息和分类列）直接进行分类，由于数据集结果只有两种类型，所以首先进行分两类的测试，结果如下：总体将 683 条数据分成了两类，总体的正确率为 94.44%，其中第一类的正确率为 93.56%，第二类的正确率为 96.31%。下面是分类后对按照不同属性绘制的属性值分布图，如图 2-25 和图 2-26 所示。

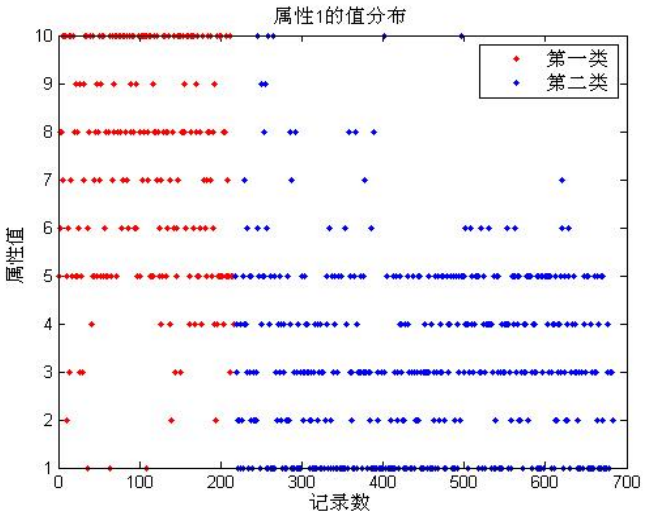


图 2-25　属性值分布图（属性 1）

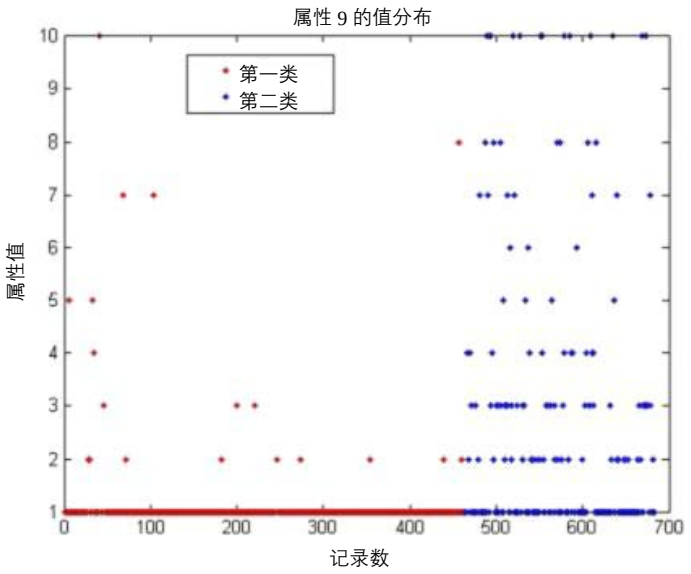


图 2-26　属性值分布图（属性 9）

如图 2-25 和图 2-26 所示，限于篇幅，只选择了上述 3 个特征属性进行图像绘制，从结果来看，可以很直观地观察到 K-Means 算法分类后的情况，第一类与第二类的分类界限比较清晰。但是不容易观察到正确和错误的情况。表 2-30 是分类结果中各个属性的聚类中心。

表 2-30　K-Means 分类的聚类中心

	第一类	第二类
属性 1	3	8
属性 2	1	7
属性 3	1	7
属性 4	1	6
属性 5	2	5
属性 6	1	10
属性 7	2	7
属性 8	1	6
属性 9	1	1

从 K-Means 算法的效果来看，能够很准确地将数据集进行分类。一方面是由于该数据集，可能是该案例特征比较明显，另一方面是由于 K-Means 算法对这种两类的作用较大。

② k-Means 结合 ReliefF 分析数据集

单从分类正确率和结果方面来看，K-Means 算法已经完全可以对乳腺癌数据集做出非常准确的判断。但是考虑 ReliefF 算法对属性权重的影响，本小节将结合 ReliefF 算法和 K-Means 算法来对该数据集进行分析，一方面得到处理该问题一些简单的结论，另外一方面可以得到一些对医学处理数据的研究方法。

首先，本小节根据第 3 小节中的一些结论，根据不同属性的权重来对 K-Means 分类数据进行预处理，以得到更精确的结论和对该数据更深度的特征规律。

从第 3 小节中，得知属性 9<属性 5<属性 7<属性 4<属性 2<属性 3<属性 8<属性 1<属性 6，根据 ReliefF 算法原理本文可以认为，对于属性 6 和属性 1 的这种重要特征属性，应该对分类起到更加大的作用。所以下面将单独对各个属性的数据进行分类测试，详细结果如表 2-31 所示。

表 2-31　单独对每个属性进行聚类分析的正确率

	总的分类正确率	第一类分类正确率	第二类分类正确率
属性 1	0.8551	0.8236	0.9667
属性 2	0.8536	0.8197	0.9793
属性 3	0.8975	0.8740	0.9617
属性 4	0.8331	0.8044	0.9433
属性 5	0.8785	0.8676	0.9063
属性 6	0.8873	0.8605	0.9655
属性 7	0.8712	0.8490	0.9364
属性 8	0.8448	0.8201	0.9290
属性 9	0.7599	0.7341	0.9412

总的分类正确率中，属性 9 最低，属性 6 最高，这与 ReliefF 算法测试的结果大致相似，但是由于 ReliefFar 算法中间部分权重接近，所以也区分不明显。说明特征属性权重的判断对分类是有影响的。上述单独分类中，只将需要分类的列数据取出来，输入到 K-Means 算法中即可。由于输入数据的变化，K-Means 分类时结果肯定是有差距的，所以单独从一个属性判断其类型是不可靠的。下面选择了单个分类时最高和最低的情况，绘制其分类属性值分布图，如图 2-27 所示。

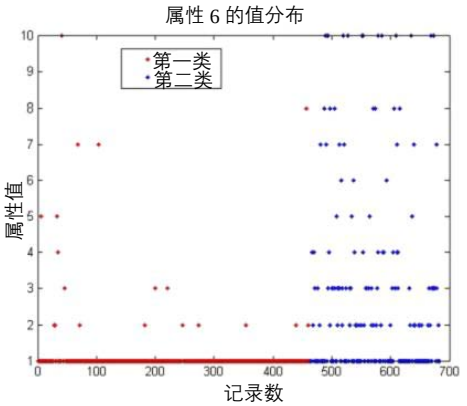


图 2-27　属性 6 的值分布

下面将对特征权重按照从大到小的顺序，选择相应的数据，进行聚类分析，结论如下：

a. 直接选择全部 9 种属性，分类成功率为：94.44%；

b. 选择属性 6，属性 1，分类成功率为：91.36%；

c. 选择属性 6、1、8、3，分类成功率为：93.85%；

d. 选择属性 6、1、8、3、2、4，分类成功率为：94.48%；

e. 选择属性 6、1、8、3、2、4、5、7，分类成功率为：95.02%。

从上面的测试可以看出，选择特征权重最大的 6 个属性，其正确率就达到选择所有属性的情况，因此我们可以认为特征权重最小的几个属性在乳腺癌诊断过程的实际作用可能比较小，实际有可能造成反作用，也就是这几个属性值与乳腺癌没有必然的联系。这一点可以作为诊断参考，或者引起注意，进行进一步的研究和确认。

③ K-Means 分成 3 类的情况

虽然从上述第 2 小节的实验中可以得到该数据集的大部分结果和结论。但是为了将相同类型的数据更加准确地分出，下面将尝试分为 3 类的情况。一方面，可以分析在乳腺癌良性和恶性情况下的显著特征属性；另一方面也可以根据此结果找到更加合理的解决方法。

还是采用 Matlab 中的 kmeans 函数，将分类数改为 3，由于分为 3 类后数据类型增多，判断较复杂，所以手动对数据进行分析，将所有特征属性加入进去。运行结果如下，测试数据中总共 683 条，其中良性共 444 条，恶性共 239 条：

a. 分为第一类的记录中，良性占 96.88%；

b. 分为第二类的记录中，恶性占 100%；

c. 分为第三类的记录中，恶性占 92%。

根据上述结果可以认为第一类为良性的分类，第二类为恶性分类，第三类为混合类。对于混合类，说明里面的数据较其他数据更加接近于偏离病例的典型数据，所以进一步分析在第一类中和第二类中的分类正确率：

a. 第一类为良性，共 448 条数据，分类正确率为 96.88%；

b. 第二类为恶性，共 99 条数据，分类正确率为 100%；

c. 第三类为混合类，共 136 条数据。

因此单独从分类后的正确率来看，效果有提高，说明对典型的病例数据分类更准确，但是对于第三类数据，而无法区分，因此这种情况下，其意义不在于分类的整体正确率，而在于在一些特殊情况下，可以根据一些重要的特征属性值就可以为患者确诊，从而提高效率和准确率，减少误诊断的几率。

上面是将所有属性进行 K-Means 变换，下面将结合 ReliefF 算法，先去掉一部分特征权重较小的特征属性后，再进行 K-Means 处理。根据第（3）小节中的结论，下面提取权重最大的 6 个属性进行测试，分别是：属性 6、属性 1、属性 8、属性 3、属性 2、属性 4。

a. 第一类为良性，共 281 条数据，分类正确率为 97.51%；

b. 第二类为恶性，共 211 条数据，分类正确率为 97.16%；

c. 第三类为混合类，共 191 条数据。

因此，对比可以看到，虽然良性的正确率增加了，但是检测出的数据减少了。第三类混合的数量也增多了，说明剔除了特种属性较小的属性，可以更加容易区分极端的病例数据，对极端数据的检测更加准确。

④ 主要的 Matlab 源代码

a. ReliefF 特征提取算法 Matlab 主程序

```
%主函数
clear;clc;
load('matlab.mat')
D=data(:,2:size(data,2));%
m =80 ;%抽样次数
k = 8;
N=20;%运行次数
for i =1:N
   W(i,:) = ReliefF (D,m,k) ;
end
for i = 1:N    %将每次计算的权重进行绘图,绘图 N 次，看整体效果
   plot(1:size(W,2),W(i,:));
   hold on ;
end
```

```
    for i = 1:size(W,2)  %计算 N 次中，每个属性的平均值
       result(1,i) = sum(W(:,i))/size(W,1) ;
    end
    xlabel('属性编号');
    ylabel('特征权重');
    title('ReliefF 算法计算乳腺癌数据的特征权重');
    axis([1 10 0 0.3])
    %------- 绘制每一种的属性变化趋势
    xlabel('计算次数');
    ylabel('特征权重');
    name =char('块厚度','细胞大小均匀性','细胞形态均匀性','边缘粘附力','单上皮细胞尺寸','裸核','Bland 染色质','正常核仁','核分裂');
    name=cellstr(name);

    for i = 1:size(W,2)
       figure
       plot(1:size(W,1),W(:,i));
       xlabel('计算次数') ;
       ylabel('特征权重') ;
       title([char(name(i))  '(属性' num2Str(i) ')的特征权重变化']);
    end
```

b. ReliefF 函数程序

```
%Relief 函数实现
%D 为输入的训练集合,输入集合去掉身份信息项目;k 为最近邻样本个数
function W = ReliefF (D,m,k)
Rows = size(D,1) ;%样本个数
Cols = size(D,2) ;%特征熟练,不包括分类列
type2 = sum((D(:,Cols)==2))/Rows ;
type4 = sum((D(:,Cols)==4))/Rows ;
%先将数据集分为 2 类，可以加快计算速度
D1 = zeros(0,Cols) ;%第一类
D2 = zeros(0,Cols) ;%第二类
for i = 1:Rows
   if D(i,Cols)==2
      D1(size(D1,1)+1,:) = D(i,:) ;
   elseif D(i,Cols)==4
      D2(size(D2,1)+1,:) = D(i,:) ;
   end
end
```

```
    W =zeros(1,Cols-1) ;%初始化特征权重，置 0
    for i = 1 : m  %进行 m 次循环选择操作
       %从 D 中随机选择一个样本 R
        [R,Dh,Dm] = GetRandSamples(D,D1,D2,k) ;
        %更新特征权重值
        for j = 1:length(W) %每个特征累计一次，循环
            W(1,j)=W(1,j)-sum(Dh(:,j))/(k*m)+sum(Dm(:,j))/(k*m) ;%按照公
式更新权重
        end
    end
```

ReliefF 辅助函数，寻找最近的样本数 K。

```
%获取随机 R 以及找出邻近样本
%D：训练集;D1：类别 1 数据集;D2：类别 2 数据集;
%Dh：与 R 同类相邻的样本距离;Dm：与 R 不同类的相邻样本距离
function [R,Dh,Dm] = GetRandSamples(D,D1,D2,k)
%先产生一个随机数，确定选定的样本 R
r = ceil(1 + (size(D,1)-1)*rand) ;
R=D(r,:); %将第 r 行选中，赋值给 R
d1 = zeros(1,0) ;%先置 0,d1 是与 R 的距离，是不是同类在下面判断
d2 = zeros(1,0) ;%先置 0,d2 是与 R 的距离
%D1,D2 是先传入的参数，在 ReliefF 函数中已经分类好了
for i =1:size(D1,1)   %计算 R 与 D1 的距离
    d1(1,i) = Distance(R,D1(i,:)) ;
end
for j = 1:size(D2,1)%计算 R 与 D2 的距离
    d2(1,j) = Distance(R,D2(j,:)) ;
end
[v1,L1] = sort(d1) ;%d1 排序,
[v2,L2] = sort(d2) ;%d2 排序
if R(1,size(R,2))==2  %如果 R 样本=2，是良性
    H = D1(L1(1,2:k+1),:) ; %L1 中是与 R 最近的距离的编号，赋值给 H。
    M = D2(L2(1,1:k),:) ; %v2(1,1:k) ;
else
    H = D1(L1(1,1:k),:);
    M = D2(L2(1,2:k+1),:) ;
end
%循环计算每 2 个样本特征之间的特征距离：(特征 1-特征 2)/(max-min)
for i = 1:size(H,1)
```

```
    for j =1 :size(H,2)
         Dh(i,j) = abs(H(i,j)-R(1,j))/9 ; % 本文数据范围都是 1-10，所以 max-min=9 为固定
        Dm(i,j) = abs(M(i,j)-R(1,j))/9 ;
    end
end
```

c. K-Means 算法主程序

```
clc;clear;
load('matlab.mat')%加载测试数据
N0 =1 ;  %从多少列开始的数据进行预测分类
N1 = size(data,1);%所有数据的行数
data=data(N0:N1,:);%只选取需要测试的数据
data1=data(:,[2,3,4,5,6,7,8,9]);% [2,4,7,9]  2:size(data,2)-1
opts = statset('Display','final');%控制选项
[idx,ctrs,result,D] = kmeans(data1,2,... %data1 为要分类的数据,2 为分类的类别数,本文只有 2 类
                    'Distance','city',... %选择的距离的计算方式
                    'Options',opts);    % 控制选项,参考 matlab 帮助
t=[data(:,size(data,2)),idx(:,1)];%把测试数据最后一列，也就是分类属性 和分类结果取出来：列 + 列
d2 = data(idx==1,11);%提取原始数据中属于第 1 类的数据的最后一列
a = sum(d2==2) ;
b=a/length(d2) ;
totalSum = 0 ;%总的正确率
rate1 = 0 ;%第一类的判断正确率.分类类别中数据的正确性
rate2 = 0 ;%第二类的判断正确率.
if(b>0.5) %说明第 1 类属于良性,则 a 的值就是良性中判断正确的个数
    totalSum = totalSum + a ;
    rate1 = a/length(d2) ;
    %然后加上恶性中判断正确的比例
    totalSum = totalSum + sum(data(idx==2,11)==4) ;
    rate2 = sum(data(idx==2,11)==4)/length(data(idx==2,11)) ;
else  %说明第 1 类属于恶性
     totalSum = totalSum + sum(data(idx==1,11)==4) ;
     totalSum = totalSum + sum(data(idx==2,11)==2) ;
     rate1 =  sum(data(idx==2,11)==2)/length(data(idx==2,11)) ;
     rate2 =  sum(data(idx==1,11)==4)/length(data(idx==1,11)) ;
end
 x1 =1;%第 x1 个属性
```

```
    x2 =1 ;%第 x2 个属性
    plot(1:sum(idx==1),data1(idx==1,x1),'r.','MarkerSize',12);
    hold on ;

plot(sum(idx==1)+1:sum(idx==1)+sum(idx==2),data1(idx==2,x1),'b.','MarkerSize
',12);
    xlabel('记录数');
    ylabel('属性值');
    title('属性 9 的值分布');
    legend('第一类','第二类');
    axis([0 640 0 10])
    rate = totalSum/size(t,1)    %总的判断准确率
```

资料来源：.NET 数据挖掘与机器学习. 作者博客: http://www.cnblogs.com/asxinyu

第 3 章

临床医学与数据技术的深度融合

- 二型糖尿病与胰腺癌的故事
- Cox 回归的基本原理与应用
- 医学数据分析中的故事
- 聚类的临床医学意义
- 贝叶斯算法的应用案例

3.1 二型糖尿病与胰腺癌的故事

在探讨疾病与疾病的共生关系中，数据挖掘技术有独特的优势。现代临床医学的数据分析表明许多疾病之间存在着共生的关系，一个临床医生只有掌握必备的数据挖掘技术或医学统计学知识才能更多地发现新的疾病规律。

如图 3-1 所示，胰腺癌是现代疾病中很难早期发现的疾病之一。在大多数癌症死亡率都在下降的今天，胰腺癌的预后指标却没有显著的改善，更重要的是由于独特的解剖位置关系，胰腺癌的手术治疗效果很差且手术的并发症多，这些特点都决定了胰腺癌的诊断需要另辟蹊径。利用疾病的共生关系来解决胰腺癌的早期诊断问题就成为一个主要的思路。临床医生们这次把目光锁定在二型糖尿病身上，最主要的原因来源于临床数据挖掘的结果：二型糖尿病与多种癌症息息相关。

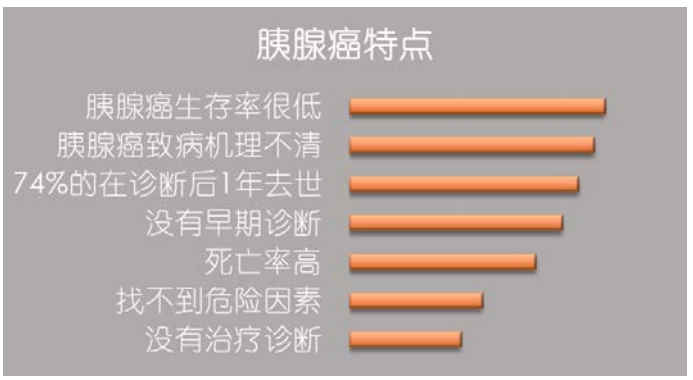


图 3-1　胰腺癌的特点

如表 3-1 所示，临床医生们选择了上述 13 个指标为变量，这些变量都是非常普通的生化指标，它们却能够帮助我们甄别二型糖尿病与胰腺癌之间的不同。

表 3-1 相关变量的选择

变 量 名 称	变 量 指 标
血糖控制和胰岛素敏感性指标	OGTT 中的空腹血糖水平(FPG) (mmol /L)
	OGTT 中的空腹胰岛素水平(FINS) (pmol /L)
	OGTT 中的餐后 2 h 血糖水平(PBG)(mmol /L)
	OGTT 中的餐后 2 h 胰岛素水平(FINS)(pmol /L)
	糖化血红蛋白(HbA1c) (%)
糖尿病相关的脂代谢指标	第一次就诊时的总胆固醇水平(TC) (mmol /L)
	第一次就诊时的甘油三酯水平(TG) (mmol /L)
	第一次就诊时的低密度脂蛋白胆固醇水平(LDL - C) (mmol /L)

续表

变量名称	变量指标
糖尿病相关的脂代谢指标	体重指数(BMI)
糖尿病相关的肝、肾功能指标	总胆红素水平(TBIL) (μmol /L)
	γ- 谷氨酰转肽酶水平(γ- GT) (u /L)
	胱抑素 C(Cystatin C) (mg /L)
	肌酐(酶法) (CR)(μmol /L)

回归系数是指在回归方程中表示自变量 *x* 对因变量 *y* 影响大小的参数。回归系数越大表示 *x* 对 *y* 影响越大，正回归系数表示 *y* 随 *x* 增大而增大，负回归系数表示 *y* 随 *x* 增大而减小。回归方程式^*Y*=*bX*+*a* 中的斜率 *b*，称为回归系数，表示 *X* 每变动 1 单位，平均而言，*Y* 将变动 *b* 单位。标准误差——统计学名词，一种量度数据分布的分散程度的标准，用以衡量数据值偏离算术平均值的程度。标准偏差越小，这些值偏离平均值就越少，反之亦然。标准偏差的大小可通过标准偏差与平均值的倍率关系来衡量。

专业上，*P* 值为结果可信程度的一个递减指标，*P* 值越大，我们越不能认为样本中变量的关联是总体中各变量关联的可靠指标。*P* 值是将观察结果认为有效即具有总体代表性的犯错概率。如 *p*=0.05 提示样本中变量关联有 5%的可能是由于偶然性造成的。即假设总体中任意变量间均无关联，我们重复类似实验，会发现约 20 个实验中有一个实验，我们所研究的变量关联将等于或强于我们的实验结果（这并不是说如果变量间存在关联，我们可得到 5%或 95%次数的相同结果，当总体中的变量存在关联，重复研究和发现关联的可能性与设计的统计学效力有关）。*t* 值其实就相当于确定了的一个置信区间，在这个区间内，接受原假设，而 *P* 表示的是置信区间之外的那部分；在确定 *t* 值时置信区间已经确定了，*P* 值也就确定了，*P* 值作为一个标准，你可以选的是显著性水平，只要比较一下就可以。两者在本质上是一样的。*t* 值是对单个变量显著性的检验，*t* 值的绝对值大于临界值说明该变量是显著的，要注意的是 *t* 检验是对总体当中变量是否是真正影响因变量的因素的检验，即检验总体中该变量的参数是否为零，只不过总体中变量的参数永远未知，只能用其无偏估量（参数的样本估计量)来代替进行检验。

如表 3-2 所示，采用 HbA1c 数据偏回归估计及假设检验我们可以看到，BMI/PBG/TC 三个变量的影响较大，DF/LDL 是负值，表明随其值增大结果反而变小。本组数据中，BMI/PBG 标准差最小，表明其与数据队列中的平均值最为靠近。在 *T* 值与 *P* 值的假设检验中除 DF 值外都有显著的统计学意义。

表 3-2　HbA1c 数据偏回归估计及假设检验

变　量	截　距	回归系数	标准误差	*T* 值	*P* 值
DF	1	–0.8652	1.9288	–0.4673	0.5856
BMI	1	0.1132	0.0564	2.3152	0.0236
PBG	1	0.2563	0.0484	5.9986	<0.0001

续表

变　量	截　距	回归系数	标准误差	*T* 值	*P* 值
TC	1	1.7453	0.5412	3.2111	0.00158
LDL	1	–1.9231	0.6639	0.6827	0.00475

从表 3-3 选取的 5 个指标来看，PC with DM（胰腺癌伴糖尿病）组与 T2DM（二型糖尿病）组差异并不十分显著，这就导致了一个问题：两个数据队列能够被显著区分吗？如果不能够被区分的话，采用共生关系来寻找胰腺癌的蛛丝马迹的企图就有可能落空。这时，数据挖掘比传统统计学的优势完全表现出来了，因为数据挖掘可以用可视化的方法把数据之间的关系表示出来，数据的可视化甚至成为数据挖掘技术的一个分支。

表 3-3　胰腺癌合并 DM 与 T2DM 比较

项　目	PC with DM	T2DM
BMI（kg/m²）	22.54 ± 2.02	23.31 ± 2.68
FPG（mmol/L）	9.62 ± 3.26	9.18 ± 3.54
PBG（mmol/L）	10.11 ± 4.02	10.72 ± 3.64
HbA1c（%）	11.23 ± 3.01	10.39 ± 2.33
FINS（pmol/L）	53.47 ± 12.24	48.69 ± 13.62

如图 3-2 所示，我们成功运用图形化曲线把两组数据队列在坐标系中有效地区分开来，在两组看似相同的数据队列中，我们用数据技术让胰腺癌的踪迹暴露出来，这就是数据技术的力量。基于大量的病例数据，我们可以从二型糖尿病的高危人群中筛查胰腺癌的高危人群。

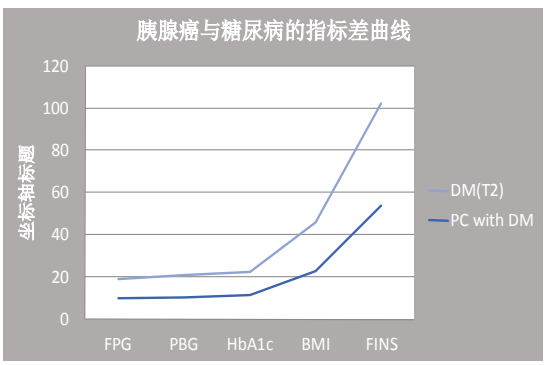


图 3-2　胰腺癌与糖尿病的指标差异曲线

如图 3-2 所示，关于胰腺癌和糖尿病的关系，至今没有明确的定论，但临床上糖尿病与胰腺癌“结伴”出现的病例又屡见不鲜。国内一些回顾性研究发现，胰腺癌患者中 4 成合并糖尿病，胰腺癌合并糖尿病患者高于非糖尿病患者。胰腺癌与糖尿病真可谓一对若即若离的“密友”。一项日本研究显示，胰腺癌病人当中之前有糖尿病的就占 12.5%，而患有胃癌、食管癌、直肠癌的病人，之前有

糖尿病的仅占 0.6%～1.2%，两者相差 10～20 倍，说明糖尿病与胰腺癌密切相关。而胰腺癌与糖尿病，看似风马牛不相及，实际上也是同出一源——胰腺。糖尿病是胰腺内分泌细胞胰岛素分泌出了问题所致，而胰腺癌多是由胰腺外分泌结构恶变而来。胰腺癌素有“癌王”之称，因为它早期缺乏显著症状，位置比较隐蔽，普通检查难以发现，所以容易被漏诊，而到了中晚期后治疗效果不理想，预后比较差。近年来糖尿病患者也是日益增多，人们对出现多食、多饮、多尿及消瘦、乏力等症状的中老年人，往往只想到二型糖尿病，而忽略了胰腺癌的可能性，以至按糖尿病治疗无效，失去了手术根治的时机。早期胰腺癌之所以会出现一些类似糖尿病的症状，是因为癌细胞破坏了胰腺组织，导致胰岛素分泌减少，因此出现高血糖和尿糖，甚至葡萄糖耐量试验也不正常。所不同的是，胰腺癌伴发的糖尿病症状按正规的降血糖治疗，难以控制，反而出现越来越严重的消化道症状。其原因有可能是因为长期的糖尿病对胰腺产生慢性刺激，使胰腺细胞易发生癌变，另外胰导管上皮细胞病理改变与胰腺癌的发生密切相关。关于糖尿病与胰腺癌，目前大概有两种说法。一种说法主张糖尿病、特别是二型糖尿病本身，就是一个诱发癌症的风险因素，它对胰岛、胰腺功能来说本身就可能是一个风险因素，所以二型糖尿病患者发生胰腺癌的风险也因此升高；而另一种说法正好相反，即认为糖尿病其实是由胰腺癌引起的，由于肿瘤影响了胰腺的功能，胰岛功能下降会导致血糖控制出现问题，也因此表现为糖尿病。糖尿病是胰腺癌的高危因素，最早有关二者关系的一篇荟萃分析显示，糖尿病患者与非糖尿病患者相比，患胰腺癌的相对风险为 2.1，其中糖尿病病史 5 年以上患者的风险为 2.0，研究者认为这一结果支持长期糖尿病作为胰腺癌的危险因素。

虽然学术界对二型糖尿病与胰腺癌的看法还没有定论，我们通过数据挖掘的办法采用五变量分析法可以将胰腺癌的高危人群从二型糖尿病患者中分离出来，做到早期诊断，早期发现。

3.2 Cox 回归的基本原理与应用

3.2.1 Cox 回归的基本原理

Cox 比例风险回归模型（Cox’s proportional hazards regression model），简称 Cox 回归模型。该模型由英国统计学家 D.R.Cox 于 1972 年提出，主要用于肿瘤和其他慢性病的预后分析，也可用于队列研究的病因探索。

Cox 回归无疑是医学数据挖掘与医学统计中最有魅力的回归分析工具，也是最常见的医学数据回归分析工具，深受广大医生的欢迎。Cox 回归是一种半参数模型，与参数模型相比，该模型不能给出各时点的风险率，但对生存时间分布无要求，可估计出各研究因素对风险率的影响，因而应用范围更广。Cox 回归是生存分析中最重要的方法之一，其优点是适用范围很广以及便于做多因素分析。

Cox 回归的因变量有些特殊，因为它的因变量必须同时有两个，一个代表状态，必须是分类变量；一个代表时间，应该是连续变量。只有同时具有这两个变量，才能用 Cox 回归分析。Cox 回归主要用于生存资料的分析，生存资料至少有两个结局变量，一是死亡状态，是活着还是死亡？二是死亡时间，如果死亡，什么时间死亡？如果活着，从开始观察到结束时有多久了？所以有了这两个

变量，就可以考虑用 Cox 回归分析。 Cox 回归与 Logistics 回归的区别如下。

① 都可以用来筛选影响因素。

② 都有 OR 值或者 RR 值。

③ 因变量不一样：Cox 回归的因变量是生存时间*Cencor（结局），而 logistic 回归因变量是分类资料，比如二分类。

④ 条件 logsitic 回归分析与 Cox 回归分析有相似的地方，SAS 程序相同，SPSS 里面条件 logistic 回归分析就是借用 Cox 比例风险模块进行分析。logistic 回归是 Cox 回归的一个特例，当全部个体都有结局时，两者的结果(β)是一样的。Cox 回归可以考察生存函数，而 logistic 不可以。在 SPSS 里，配对 logistic 回归的模型，是在 Cox 回归里完成的。

Cox 回归假定病人的风险函数如下。

$h(t/X)=h_0(t)\exp(\beta_1 X_1+\beta_2 X_2+\cdots\cdots+\beta_p X_p)$

$h_0(t)$：基准风险函数，即所有变量取零时的 t 时刻的风险函数

X_1、X_2 …… X_p：影响因素变量。

β_1、β_2 …… β_p：回归系数。

3.2.2　晚期肺癌伴脑转移患者的预后多因素 Cox 回归

案例：

（1）主要研究目标：探讨影响非小细胞肺癌（non-small cell lung cancer，NSCLC）脑转移患者生存时间的因素。

方法：回顾性分析成都市第三人民医院收治的 NSCLC 脑转移并行头颅放疗患者 302 例，其中对资料完整者 171 例进行分析。采用 SPSS 13.0 统训一软件行影响生存期的单因素及多因素 Cox 风险比例模型回归分析。探讨患者的临床特征及放疗方式等因素对患者生存期的影响。

结果：全组患者中位生存期为 8.8（95% CI：7.2-10.3）个月；单因素分析显示：PS 评分（P = 0.002）、脑转移数量（P=0.023）、脑转移时间（P=0.031）、放疗方式（P=0.041）和肺癌是否切除（P = 0.002）与患者预后有关；Cox 多元回归分析显示：PS 评分（P= 0. 04）和肺癌是否手术切除（P = 0.04）为脑转移患者独立预后因素，而与脑转移数量（P=0.65）、脑转移时间（P=0.71）、放疗方式（P=0.91）等因素无关。

结论：NSCLC 脑转移整体预后较差，手术切除肺部肿瘤且体力评分较好患者预后相对较好。

（2）晚期肺癌伴脑转移的医学统计学背景

肺癌患者确诊时 80%已为晚期，大多数数失去了手术机会。流行病学资料显示约 30%的 NSCLC 患者最终会发展为脑转移，发生脑转移的 NSCLC 患者往往预后较差，未经治疗的肺癌脑转移患者中位生存期大约为 3～6 个月。立体定向放疗（stereotactic radiotherapy SRT）或 SRT 联合全脑放疗（whole brain radiotherapy，WBRT）可以使生存期延长至 7～15 个月。但总体来讲，NSCLC 出现脑转移已为肺癌晚期，往往预后不良，只有少部分患者生存期超过 18 个月。

（3）数据收集

回顾性分析成都市第三人民医院近年来收治的 NSCLC 脑转移患者 302 例，其中对随访资料完整者 171 例进行分析。本组患者中男性 119 例（69.6%），女性 84 例（30.4%），年龄 42～79，平均年龄 57.8111 岁。其中鳞癌 51 例（29.8%）、腺癌 92 例（53.8%），其他病理类型肺癌 28 例（16.4%）；PS：小于等于 2 分 81 例（47.4%），大于 2 分 90 例（52.6%）；非吸烟者 61 例占 35.7%，吸烟者 110 例占 64.3%；行 WBRT 者 88 例（51.5%），行 SRT 者 52 例（30.4%），行 SRT+WBRT 者 31 例（18.1%）。本组患者的临床基线情况见表 3-4。

（4）治疗方法

放疗方法 105 例患者接受了 SRT 放疗，放疗计划采用 CREAT XSTPS 系统制定，瓦里安 600 C 直线加速器 6 MV X 线照射。所有患者制订放疗计划时根据组织密度进行矫正。根据国际辐射单位和测量委员会（ICRU-2）指南勾画靶区，应包括大体肿瘤靶区（GTV）、微小病灶的临床靶区（CTV）、靶区运动的内靶区（ITV）边缘和每日靶区定位误差边缘来制定的计划靶区（PTV）。多发脑转移患者给以全脑放疗和或 SRT。放化疗过程中检测患者血象变化，并给以升白细胞药物和对症支持。如患者放疗过程中出现严重的 111-W 血液学毒性或恶性呕吐等不能耐受者予以暂停放疗，待血象恢复正常或符合放疗要求后继续放疗。

随访 302 例 NSCLC 脑转移患者中随访资料完整者为 171 例，随访的终点事件为患者死亡。生存时间定义为患者确诊肺癌之日至死亡的时间，以月为单位进行计算。

（5）统计分析

患者生存时间采用中位数表示，组间比较采用 Log-rank 检验。Kaplan-Meier 风险比例模式绘制生存曲线，计算中位生存时间及 95%置信区间。影响患者生存期的多因素分析采用 Cox 回归模型进行分析。

表 3-4　单因素分析与患者预后有关的相关因素

临床特征	*n*（%）	中位生存时间（month）	Log-rank test x^2	*P*
性别			1.11	0.32
男	119（69.6%）	8.5		
女	84（30.4%）	9.36		
病理类型			0.78	0.56
鳞癌	51（29.8%）	8		
腺癌	92（53.8%）	7.6		
其他	28（16.4%）	9.2		
吸烟情况			0.71	0.61
不吸烟	61（35.7%）	8.2		
吸烟	110（64.3%）	9.0		
PS 评分			9.8	0.002
≤2 分	81（47.4%）	15.2		

续表

临床特征	*n*（%）	中位生存时间（month）	Log-rank test x^2	*P*
>2 分	90（52.6%）	5.1		
脑转移数量			5.3	0.23
单发	68（39.8%）	10.2		
多发	103（60.2%）	6.1		
脑转移时间			4.2	0.031
≤5 个月	51（29.8%）	7.1		
>5 个月	120（70.2%）	11.2		
放疗方式			3.8	0.041
WBRT	88（51.5%）	7.8		
SRT	52（30.4%）	10.6		
SRT+WBRT	31（18.1%）	8.1		
化疗			1.21	0.30
是	103（60.2%）	8.1		
否	68（39.8%）	9.3		
肺癌手术				
肺癌切除	51（29.8%）	17.5	9.6	0.002
未切除	120（70.2%）	6.1		

资料来源：骆竹梅. 非小细胞肺癌脑转移患者预后多因素 Cox 回归分析. 成都市第三人民医院. 临床肺科杂志. 2015 年 4 月.

如表 3-4 所示，对患者临床特征如性别、吸烟史、病理类型、PS 评分、脑转移数量、肺癌至脑转移时间、放疗方式等情况与预后关系进行单因素分析，结果显示：PS 评分（*P*=0.002）、脑转移数量（*P*=0.023）、脑转移时间（*P*=0.031）、放疗方式（*P*=0.041）和肺癌是否切除（*P*=0.002）与患者预后有关（表 3-4）。其中 PS 评分主要指评价患者的体力活动状态（performance status，PS），即从患者的体力来了解其一般健康状况和对治疗的耐受能力。

时序检验（log-rank test）两个生存率曲线不同，它们之间的差别有无统计学意义呢？比如，在临床实验中，对照组（A 组）和治疗组（B 组）的生存曲线之差别有无统计学意义？回答这个问题可用时序检验。时序检验是计算出不同日期两种疗法的暴露人数及死亡人数，并由此根据两种疗法疗效相同的假设计算出两种疗法在该日的期望死亡数，如无效假设是对的，则实际值与期望值不会相差很大；如相差过大则不仅仅是由于机遇所产生的差异。对此可做检验以判断。

如表 3-5 所示，PS 评分（*P*=0.002）、脑转移数量（*P*=0.023）、脑转移时间（*P*=0.031）、放疗方式（*P*=0.041）和肺癌是否切除（*P*=0.002）等与 NSCLC 脑转移患者预后有关的变量代入 Cox 回归方程进行多因素分析，结果显示：PS 评分（*P*=0.04）和肺癌是否手术切除（*P*=0.04）为脑转移患者独立预后因素，而与脑转移数量（*P*=0.65）、脑转移时间（*P*=0.71）、放疗方式（*P*=0.91）等因素

无关（表 3-5）。

表 3-5　影响患者预后因素的 Cox 回归分析

临床特征	*B*	SE	Wald	df	*P*	OR	95%CI 或 OR
PS	-0.75	0.22	7.81	1	0.04	0.66	0.32~0.89
脑转移数量	-0.21	0.05	2.12	1	0.65	0.89	0.61~2.35
脑转移时间	0.16	0.1	1.23	1	0.71	1.23	0.68~2.64
放疗方式	0.1	0.08	0.98	1	0.91	1.12	0.56~2.24
肺癌手术	0.83	0.31	8.21	1	0.04	1.66	1.20~3.62

资料来源：骆竹梅. 非小细胞肺癌脑转移患者预后多因素 Cox 回归分析. 成都市第三人民医院. 临床肺科杂志. 2015 年 4 月.

近年来我国肺癌的发病率呈逐渐上升趋势，未来几年我国将成为全球肺癌发病人数最多的国家。流行病学资料显示约 30%的 NSCLC 患者最终会发展为脑转移，发生脑转移的 NSCLC 患者往往预后较差，未经治疗的肺癌脑转移患者中位生存期大约为 3～6 个月。因此，脑转移是导致肺癌死亡的一个重要原因。WBRT +/-SRT 可以使生存期延长 4～6 个月，除放疗可以使肺癌脑转移患者生存期延长外，患者自身的临床特征同样对预后有一定的影响，如病灶的控制程度、PS 评分等，可能也是患者预后的重要影响因素。

本研究对 171 例肺癌脑转移患者的生存资料进行了回顾性分析，研究结果显示：全组 171 例肺癌脑转移患者的中位生存期为 8.8（95% CI ：7.2～10.3）个月，总体预后不良，生存期较短。单因素分析显示 PS 评分、脑转移数量、脑转移时间、放疗方式和肺癌是否切除与患者预后有关。Cox 多元回归分析显示只有 PS 评分（*P*=0.04）和肺癌是否手术切除（*P*=0.04）是影响患者独立预后的独立因素，而与脑转移数量（*P*=0.65）、脑转移时间（*P*=0.71）、放疗方式（*P*=0.91）等因素无关。研究结果提示手术切除肺部肿瘤且体力评分较好患者可能预后相对较好，与既往的研究结果相似。

PS 评分是目前评价肺癌患者机体行为状态的最常用指标，它可以较为准确地反映患者机体功能状况，晚期肺癌患者的机体功能状态与预后关系已较为明确，但 PS 评分能否用于 NSCLC 脑转移患者预后评价指标仍存在一定的争议。有研究报道 PS 评分是肺癌脑转移独立的预后因子，PS 评分较低的患者生存期较长；但亦有研究显示 PS 评分与 NSCLC 脑转移患者的预后无关。本研究结果显示：PS 较低者其中位生存期较长，该评分是脑转移患者的独立预后因素。分析 PS 评分对预后影响不一致的原因可能与纳入研究的患者间存在较大的临床异质性有关，不同研究纳入研究的肺癌患者的临床分期、治疗方案等存在较大差异，导致分析结果产生较大分歧。结果不一致的原因还可能与大多数研究为回顾性研究有关，回顾性分析往往受到多种偏倚的影响，这些偏倚的存在可能是导致结论存在差异的重要原因。颅外病灶的控制情况是影响肺癌患者生存的又一独立危险因素，本研究中肺癌原发灶被切除者中位生存期为 17.5 个月，显著高于未行手术治疗者 6.1 个月。Cox 回归分析同样显示手术切除肺部肿瘤病变患者的生存优势是非手术者的 1.66 倍。因此，原发病变被控制的脑转移患者往往预后较好。

总体来说，晚期肺癌患者尤其是出现脑转移患者的预后较差，生存期较短。放疗等治疗手段并不能大幅度延长患者的生存期。同时晚期 NSCLC 患者的预后受较多因素的影响，各研究结果间的差异较大，结论不尽一致。分析原因可能和纳入研究患者的临床异质性、回顾性分析本身的局限性、病例数少等因素有关。因此，评价晚期肺癌患者尤其是 NSCLC 脑转移患者的预后因素时应多方面综合考虑，不可依据一个因素而下结论。

资料来源：骆竹梅 “非小细胞肺癌脑转移患者预后多因素 Cox 回归分析. 成都市第三人民医院. 临床肺科杂志. 2015 年 4 月.

表 3-6　案例中表 3-4、表 3-5 的术语解释

术语缩写	术语的含义	术语的思想
B 值	*B* 值是指回归系数和截距，左边对应的是 constant（常数）则代表截距，即 $y=b+b_1x_1+b_2x_2$.....中的常数 *b*	回归系数是指在回归方程中表示自变量 *x* 对因变量 *y* 影响大小的参数。回归系数越大表示 *x* 对 *y* 影响越大，正回归系数表示 *y* 随 *x* 增大而增大，负回归系数表示 *y* 随 *x* 增大而减小
SE 值	标准误差(standard error),即样本均数的标准差,是描述均数抽样分布地离散程度及衡量均数抽样误差大小的尺度	一种量度数据分布的分散程度的标准，用以衡量数据值偏离算术平均值的程度。标准偏差越小，这些值偏离平均值就越少，反之亦然。标准偏差的大小可通过标准偏差与平均值的倍率关系来衡量
DF 值	自由度（degree of freedom, DF）在数学中能够自由取值的变量个数	在统计学中，自由度指的是计算某一统计量时，取值不受限制的变量个数。通常 df=*n*-*k*。其中 *n* 为样本含量，*k* 为被限制的条件数或变量个数，或计算某一统计量时用到其他独立统计量的个数。自由度通常用于抽样分布中
Wald 值	Wald 是一个卡方值，等于 *B* 除以它的标准误差（S.E.）的平方值。Wald 用于对 *B* 值进行检验	Wald 检验的思想是：如果约束是有效的，那么在没有约束情况下估计出来的估计量应该渐进地满足约束条件，因为 MLE 是一致的。以无约束估计量为基础可以构造一个 Wald 统计量，这个统计量也服从卡方分布
P 值	*P* 值即概率，反映某一事件发生的可能性大小	统计学根据显著性检验方法所得到的 *P* 值，一般以 $P < 0.05$ 为显著，$P<0.01$ 为非常显著，其含义是样本间的差异由抽样误差所致的概率小于 0.05 或 0.01。实际上，*P* 值不能赋予数据任何重要性，只能说明某事件发生的几率
OR 值	OR 是统计学中优势比的简称	在病例—对照研究中 OR 指病例组暴露人数与非暴露人数的比值(a/b)除以对照组暴露人数与非暴露人数的比值(*c*/*d*)，即 *ad*/*bc*。如果疾病的发病率很低（例如观察期间的累计发病率为 2%），且所得病例为无选择偏倚的新发病例，则 OR 为相对危险度 RR 的近似估计值。在定群研究或横断面研究中 OR 指的是暴露组中患者与非患者的比值(*a*/*c*)与非暴露组中患者与非患者的比值(*b*/*d*)之比，即 *ad*/*bc*，实际上是暴露状况的比值比。由定群研究资料计算的 OR 是相对危险度 RR 的估计值。在横断面研究中计算的 OR 是研究患者病例而不是新发生病例所得出的比值比。上述的 OR 都应进行假设检验及可信区间的计算

续表

术语缩写	术语的含义	术语的思想
CI 值	置信区间（CI）即按一定的概率估计总体参数所在的范围。	进行假设检验，95%的 CI 与 a=0.05 的假设检验等价。当效应值是比值时，若 95%的 CI 包含了 1，等价于 P>0.05，无统计学意义。当效应值是差值时，若 95%的 CI 包含了 0，等价于 P>0.05，无统计学意义
Log-rank test	时序检验(log-rank test)	生存率计算采用 Kaplan-Meier 方法，用时序检验(log-rank test)进行生存率的单因素比较，用比例风险模型进行多因素分析，采用 AIC 值评价模型预后价值的贡献大小

3.2.3 本案例的几点启示

① Cox 多因素回归模型在临床的应用中的确有不可替代的作用，Cox 回归用于生存分析数据，这一点所有人都知道。而且从现实应用情况来看，大多数人做生存分析的多因素分析时，都会选择 Cox 回归，而不是其他回归（如 Weibull 回归、指数回归等）。这是因为 Cox 回归在分析时无须考虑数据分布，直接便可以应用。而其他生存分析的回归方法，如 Weibull 回归、指数回归，需要先看一下数据是否符合相应分布，只有符合，才能用相应的回归方法，否则便不能用。所以从简便的角度出发，更多人喜欢用 Cox 回归。

② Cox 回归跟 Logistic 回归模型十分相近，极尽简单和优美的特色，它通过死亡风险比 HR 来反映研究因素的危险大小，将理论模型与实际解释联系起来，因而广受流行病学工作者的欢迎。Cox 回归和 logistic 回归是流行病学工作中两大主要数据分析工具。两种模型都可以通过对回归系数取指数，即 $EXP(\beta_i)$，表示某因素的危险大小，Logistic 回归中表示事件发生的危险大小，而 Cox 回归中表示死亡风险大小。

③ Cox 回归尽管应用广泛，但也不是说任何生存数据都可以用它来分析。它有一个重要的前提假设，即等比例风险（Porportional hazards），它表示某因素对生存的影响在任何时间都是相同的，不随时间的变化而变化。

④ 在晚期癌症的诊疗方案中，国内医学文献对数据挖掘工具的利用效率还十分落后，其中对大样本、规范性数据的获取仍然是制约中国临床医生科研水平的重要因素，临床数据标准不规范，病例数据不能共享是制约中国临床科研水平的首要原因。

⑤ 虽然 171 例完整病例的分析从统计学意义上讲有理论的样本价值，可以在一定程度上反映数据全集的情况，但由于抽样的数据完全依靠病例的搜集，因而在抽样调查的随机性原则、均匀分布原则、分层抽样原则的约束下数据的真实性、完整性大打折扣，样本的代表性有理由受到质疑，这也是目前我国医学论文普遍面临的问题。

⑥ 对晚期肺癌伴脑转移灶患者的预后回归是一项极有意义的工作。长期以来面临手术、放化疗、靶向治疗的诸多争议，在数据技术日益发展的今天，用大数据与数据挖掘的办法原本可以有效地回答这个问题，遗憾的是由于病例数据不能有效共享，临床医生们找不到大样本的数据来分析晚期肺癌伴脑转移灶患者的预后数据来有力证明最佳的治疗方案，各种文献对此也语焉不详，于是乎只能

出现本案例中的“NSCLC 脑转移整体预后较差，手术切除肺部肿瘤且体力评分较好患者预后相对较好。”这样的简单化结论。

小知识：卡方检验

卡方检验是用途非常广的一种假设检验方法，它在分类资料统计推断中的应用包括：两个率或两个构成比比较的卡方检验、多个率或多个构成比比较的卡方检验以及分类资料的相关分析等。

卡方检验就是统计样本的实际观测值与理论推断值之间的偏离程度，实际观测值与理论推断值之间的偏离程度就决定卡方值的大小，卡方值越大，越不符合，偏差越小，卡方值就越小，越趋于符合，若量值完全相等时，卡方值就为 0，表明理论值完全符合。

（1）提出原假设：

H_0：总体 X 的分布函数为 $F(x)$

如果总体分布为离散型，则假设具体为

H_0：总体 X 的分布律为 $P\{X=x_i\}=p_i,\ i=1,2,\ldots$

（2）将总体 X 的取值范围分成 k 个互不相交的小区间 A_1、A_2、A_3、…、A_k，如可取 $A_1=(a_0, a_1]$，$A_2=(a_1,a_2]$，…，$A_k=(a_k\text{-}1,a_k)$。

其中 a_0 可取-∞，a_k 可取+∞，区间的划分视具体情况而定，但要使每个小区间所含的样本值个数不小于 5，而区间个数 k 不要太大也不要太小。

（3）把落入第 i 个小区间的 A_i 的样本值的个数记作 f_i，成为**组频数（真实值）**，所有组频数之和 $f_1+f_2+\ldots+f_k$ 等于样本容量 n。

（4）当 H_0 为真时，根据所假设的总体理论分布，可算出总体 X 的值落入第 i 个小区间 Ai 的概率 p_i，于是，np_i 就是落入第 i 个小区间 A_i 的样本值的**理论频数（理论值）**。

（5）当 H_0 为真时，n 次试验中样本值落入第 i 个小区间 A_i 的频率 f_i/n 与概率 p_i 应很接近，当 H_0 不真时，则 f_i/n 与 p_i 相差很大。基于这种思想，皮尔逊引进如下检验统计量

$X^2=\sum_{i=1}^{k}\frac{(f^i-np^i)^2}{np^i}$，在 0 假设成立的情况下服从自由度为 k-1 的卡方分布。

3.3 医学数据分析中的故事

故事一：

1992 年，抗抑郁药物帕罗西汀（Paxil）获准上市；1996 年，降胆固醇药物普拉固（Pravachol）正式开售。两种药品生产企业的研究证明：每种药物在单独服用时是有效且安全的。可是，患者要是同时服用两种药是否安全，没有人知道，甚至很少有人想过。美国斯坦福大学的研究人员应用数据挖掘技术分析了数万例患者的电子病历后，很快发现了一个出人意料的答案：同时服用两种药物的患者血糖含量较高。这对于糖尿病患者来说影响很大，过多的血糖对他们来说是一种严重的健康威胁。科学家还通过分析血糖检测结果和药物处方，来寻找隐藏的规律。

对于单个医生来说，他所经历的同时服用这两种药物的病人是很有限的，虽然其中可能有少数的糖尿病患者莫名其妙地血糖升高了,但医生很难意识到这是由于病人同时服用了 Paxil 和 Pravachol 造成的。因为这是一种掩藏在大数据中的隐含规律，如果不是有人有目的地专门研究 Paxil 和 Pravachol 联合用药的安全性的话，个体医生是很难揭示这个规律的。但是，临床药品成千上万，我们怎么可能对任意组合的两、三种药联合应用的安全性和有效性进行逐一研究呢？数据挖掘很可能是一种有效的、快速的、主动式的探索多种药联合应用问题的方法！

研究者不必再召集患者去做临床试验，那样做的话花费太大了。电子病历及其计算机应用的普及为医疗数据挖掘提供了新的机遇。科学家不再局限于通过召集志愿者来开展传统的课题研究，而是更多地从现实生活中的实验中，如日常的大量的临床案例中，筛选数据并开展虚拟科研，这些并非来自计划的课题立项的实验数据，保存在许多医院的医疗记录中。

类似本案例，应用数据技术使得研究人员可以找出在药物批准上市时无法预见的问题，例如一种药物可能对特定人群产生怎样的影响。另外，对医疗记录的数据挖掘不仅将为研究带来好处，还会提高医疗服务系统的效率。

故事二：

上海的柴先生今年 55 岁，两年前因为冠心病放了 4 个心脏支架，之后一直服用**阿司匹林**、**氯吡格雷**、**美托洛尔**、**他汀**、**欣康**这 5 种药。柴先生听说放支架后氯吡格雷服用 1 年就可停掉，但是咨询了医生说法不一，柴先生很疑惑放支架后抗凝药到底要吃多久，而且长期吃会不会有药物副作用。冠心病其实就是冠状动脉狭窄或闭塞，引起闭塞血管远端的心肌缺血。所以，冠心病发作就要马上“打通”血管，改善心肌缺血。而“打通”血管有两个办法，就是不少患者都听过的“心脏搭桥”和“心脏支架手术”。不过，放了支架不等于后顾无忧，由于发展到冠心病的病人身体仍然存在多种危险因素，所以术后需要服用一些药物。

无独有偶，柴先生同时还患有糖尿病，经常需要服用**格列美脲降**糖药。治疗糖尿病的药物很多：① 磺脲类胰岛素促泌剂：如格列齐特、格列美脲；② 格列奈类促泌剂：如瑞格列奈；③ 双胍类：如二甲双胍；④ 糖苷酶抑制剂：如阿卡波糖；⑤ 胰岛素增敏剂：如吡格列酮；⑥ DPP-4 抑制剂：如沙格列汀；⑦ GLP-1 受体激动剂：如利拉鲁肽等。

突然有一天，柴先生因为低血糖晕倒在家中，后来送医院急救才恢复过来。这一现象引起了研究人员的注意，因为已经有多起上述 6 种药一起吃而快速导致低血糖的案例。研究人员用 SQL 语句调取了上海 5 家三甲医院心内科行支架手术并糖尿病的患者、医嘱用药中 6 种药物一起用的患者并大量的随访资料数据，最后发现 6 种药物长期同时服用会导致低血糖的突然出现，病人马上晕倒，生命垂危。这又是一个很好的数据挖掘案例，如果没有大数据佐证 6 种药物的关联性，也许这一现象会被当成偶然事件而忽略，正是因为数据技术的出现，一个药物冲突的事件真相才被还原与发现。

3.4 聚类的临床医学意义

3.4.1 聚类算法的基本定义

聚类又称群分析，它是研究（样品或指标）分类问题的一种统计分析方法，同时也是数据挖掘的一个重要算法。

聚类（Cluster）分析是由若干模式（Pattern）组成的，通常，模式是一个度量（Measurement）的向量，或者是多维空间中的一个点。

聚类分析以相似性为基础，在一个聚类中的模式之间比不在同一聚类中的模式之间具有更多的相似性。聚类起源于分类学，在古老的分类学中，人们主要依靠经验和专业知识来实现分类，很少利用数学工具进行定量的分类。随着人类科学技术的发展，对分类的要求越来越高，以致有时仅凭经验和专业知识难以确切地进行分类，于是人们逐渐地把数学工具引用到了分类学中，形成了数值分类学，之后又将多元分析的技术引入到数值分类学形成了聚类分析。聚类分析内容非常丰富，有系统聚类法、有序样品聚类法、动态聚类法、模糊聚类法、图论聚类法、聚类预报法等。

现实医学的应用可能需要在各种约束条件下进行聚类。假设你的研究是在一个队列中为给定数目的各种病例进行聚类，为了做出决定，你可以对病种的分子特性进行聚类，同时考虑如病理学的分子检测，对免疫细胞特性描述要求等情况。要找到既满足特定的约束，又具有良好聚类特性的数据分组是一项具有挑战性的任务。

如图 3-3 所示，在数据挖掘的实践中，聚类的算法分为划分法（K-Means 算法，K-MEDOIDS 算法、CLARANS 算法）、层次法（BIRCH 算法、CURE 算法、CHAMELEON 算法）、密度算法（DBSCAN 算法、OPTICS 算法、DENCLUE 算法等）、图论聚类法、网格算法（STING 算法、CLIQUE 算法、WAVE-CLUSTER 算法）、模型算法。

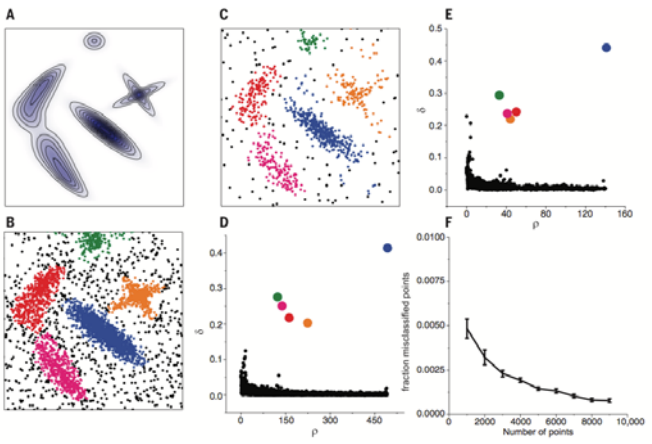


图 3-3　几种聚类算法的可视化演示

3.4.2 临床医学数据挖掘中聚类的意义

下面用一个医学数据挖掘的聚类实例来说明聚类方法的临床医学意义。

案例：基于聚类的心电信号分类方法研究

（1）心电信号背景知识

心电信号是心脏活动的一种客观表现方式，是一种典型的生物信号，具有频率、振幅、相位、时间差等特征要素，如图 3-4 所示。由于心电信号从不同方面和层次上反映了心脏的工作状态，因此心电检测系统在心脏疾病的临床诊断和治疗过程中具有非常重要的参考价值。心电图（Electrocardiogram，ECG）指的是在心脏的每个跳动周期中，由起搏点、心房、心室相继兴奋，伴随着心电图生物电的变化，通过心电描记器从体表引出多种形式的电位变化的图形，由“心电图之父”荷兰教授 Einthoven 在 1903 年发明。心电图能准确地反映出心脏兴奋的电活动过程，它对心脏基本功能及其病理研究方面，具有重要的参考价值，常用于对各种心律失常、心室心房肥大、心肌梗死、心率失常、心肌缺血、电解质紊乱、心衰等病症的检查，也可用于床边 24 小时昼夜监视病人心脏。

图 3-4　心电脉冲信号的样例

心电信号的脉冲数据解读是一个专业性很强的工作，往往受到医生个人知识水平、经验的限制，因而十分适合数据挖掘与机器学习的方法来实现大规模数据的批量处理。基本的技术路线图如图 3-5 所示。

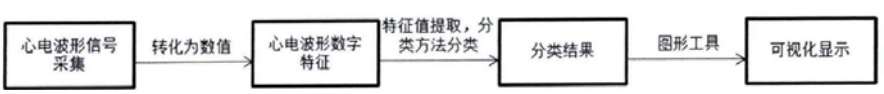


图 3-5　心电信号数据挖掘的聚类技术路线图

（2）技术路线

从图 3-5 不难看到，首先将 MIT-BIH 数据库中的 11 万份心电图信号作为实验数据来做分析和研

究，着重针对 QRS 波群进行实验，实验过程主要可分为心电信号波形数据特征提取、使用聚类算法分类和对分类结果优化 3 个部分。MIT-BIH 数据库是麻省理工学院创建的全球最大的心律失常数据库。

如图 3-6 所示，QRS 波是指正常心电图中幅度最大的波群，反映心室除极的全过程。正常心室除极始于室间隔中部，自左向右方向除极，故 QRS 波群先呈现一个小向下的 q 波。正常胸导联 QRS 波群形态较恒定。V1、V2 导联多呈 rS 型，R/S<1.Rv1.<1.0mV，超过此值常提示右心室肥大。V5、V6 导联以 R 波为主，R/S>1.Rv.<2.5 mV，超过此值常提示左心室肥大。V3、V4 导联呈 RS 型，R/S 接近于 1，称为过渡区图形。正常成人胸导联自 V1 至 V6 R 波逐渐增大，而 S 波逐渐变小。若过渡区图形（RS 型）出现于 V5、V6 导联，且 R/S 比例仍向右递减，提示心脏沿长轴发生顺钟向转位（从心尖往上看），此时右心室向前、向左旋转；若过渡区图形出现于 V1、V2 导联，且 R/S 比例仍向左递增，提示心脏沿长轴发生逆钟向转位，此时左心室向前、向右旋转。顺钟向转位可见于右心室肥大，逆钟向转位可见于左心室肥大。但这种转位图形亦可见于正常人。

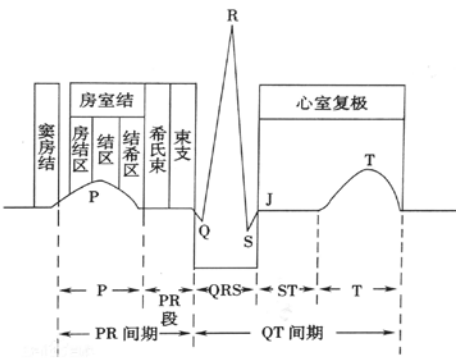


图 3-6　QRS 波型图

（3）聚类的原理

按照相似的特征进行归类，聚类分析就是根据模式的特征空间分布，按照点与点之间的距离大小来确定它们的相似度。聚类分析起源于分类学，但是聚类又不等于分类。聚类与分类的不同在于，聚类所要求划分的类是未知的，它不需要先把各类抽样组成训练集，按照训练集的统计参数去构成分类器，所以聚类方法也被称为无监督法。对数据进行聚类有以下要点：为了将样本分成多个类别，首先要确定如何度量样本间的相似性，其中一种度量方法就是样本在特征空间的距离，一个合理的相似性度量距离可以使得样本之间在同一类的距离明显小于与在非同类样本之间的距离。

（4）心电信号的特征提取

① 根据模式识别的理论，有了采样数据后需要对数据进行预处理。从 MIT-BIH 数据库中下载下来的心电信号数据是以二进制方式存储的，我们将其按照存储格式转换成十进制。我们根据 MIT-BIH 中标注出的波形的 R 点（在图 3-6 中用圆点标示）进行数据切割，得到的最终数据中一条心电信号记录为 R 点前 50 点，后 49 点，总计 100 点数据，如图 3-7 所示。

图 3-7　R 点前后 50 点的拼图

资料来源：马国伟. 基于聚类的心电信号分类方法研究. 华东理工大学 2012 年.

② 数据的降维处理：在图 3-7 中各个波形通过首尾相连拼接在一起。事实上，我们可以把这由 100 个数据点组成的波形直接用来分类，然而 100 维的数据对于作为输入到分类器中的特征值来说显得过于冗余，左右两侧的数据对于心电信号主要形态来说意义不大，在医学上也主要是根据波形的 P 波、QRS 波群来进来病症的诊断，而且 100 维的数据对于计算机计算和硬件支持也带来不少压力，因此需要对数据进行有效特征值提取或降维。从研究意义上来说，QRS 波群对于心电信号来说最为重要，从形态学上分析 QRS 波群可以诊断出许多心脏疾病，故本文中也将 QRS 波群作为主要的研究对象。根据对心电图的介绍，可以得知 QRS 波群持续的总时间不超过 0.1 秒，正常人在 0.06～0.10 秒之间，对于非正常情况下可能会有所增加。MIT-BIN 数据库的数据采样率为 360Hz，即 1 秒有 360 个数据点，通过这个我们可以大致计算出一个 QRS 波群约有 36 个点组成。考虑到实际情况下时间可能有所偏差，且 R 点不一定在 QRS 波群的中心，加上 P 波对诊断病症也有一部分作用，于是稍微增加了取的数据点的个数，分别是 R 点前 30 点，R 点，R 点后 24 点，总计 55 个数据点。图 3-8 是经过截取后的波形图，从图中我们可以看到，55 个数据点仍然能很好地表现出一个心电波形的大致形态，特别是对于一个正常的 QRS 波群来说，分布在其上的点并不多，因此 55 个数据点来表示一个心电信号波形是可行的。

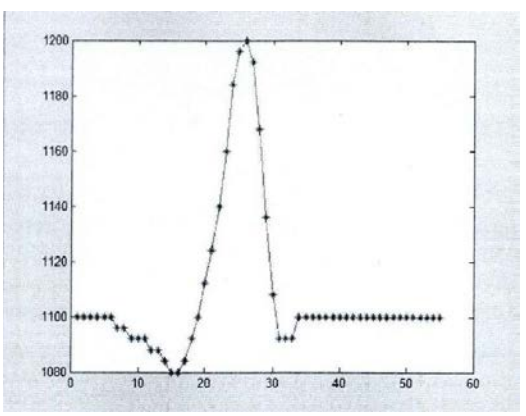


图 3-8　一个截取后的 QRS 波

资料来源：马国伟. 华东理工大学. 2012 年.

③ 曲线拟合

对数据进行特征值提取的方法有很多，如小波分析、神经网络训练等，曲线拟合也是其中一种很常用的方法。所谓的曲线拟合（Curve fitting），就是推求一个解析函数 $Y=f(x)$，使其通过或近似地通过有限序列的数据点(x,y)。形象地说，就是使用一条光滑的曲线近似地去逼近一个平面上的一系列点，这是一种用解析式逼近离散数据的方法。在求得一个解析式后，就等于把离散数据点转化为函数中的参数表示，这样能起到有效的降维作用。通常使用最小二乘法来求解函数中的待定参数。根据拟合所用的函数的不同，拟合的名称也有所不同，常用的拟合函数有指数函数、对数函数、幂函数等。在用一个解析式表示一组序列数据点后，需要有一些数学指标去评价拟合结果的好坏。

④ 傅里叶级数拟合

对于曲线拟合来说，选取一个和原始数据大体上较相似的曲线来逼近对于最终的拟合效果十分重要。对于心电信号波形来说，光是使用一般的简单函数，如幂函数、对数函数等，显然是不合适的。如图 3-9 所示，笔者尝试选用傅里叶级数（Fourier series）作为曲线拟合函数。傅里叶级数由法国数学家傅里叶发现，他指出任何周期函数都可以用正弦函数和余弦函数构成的无穷级数来表示。傅里叶级数在数论、信号处理、统计学、声学等领域都有广泛应用。傅里叶变换就是将满足某个条件的函数表示成三角函数或者它们的积分的线性组合。在信号处理中，傅里叶变换的经典用途就是把信号分解成振幅分量和频率分量。在不同的研究领域，傅里叶变换具有多种不同的变体形式，如连续傅里叶变换和离散傅里叶变换，离散形式的傅里叶变换可以利用数字计算机快速实现。连续形式的傅里叶变换其实是傅里叶级数的推广，因为积分其实是一种极限形式的求和算子而已。对于周期函数，其傅里叶级数是存在的。

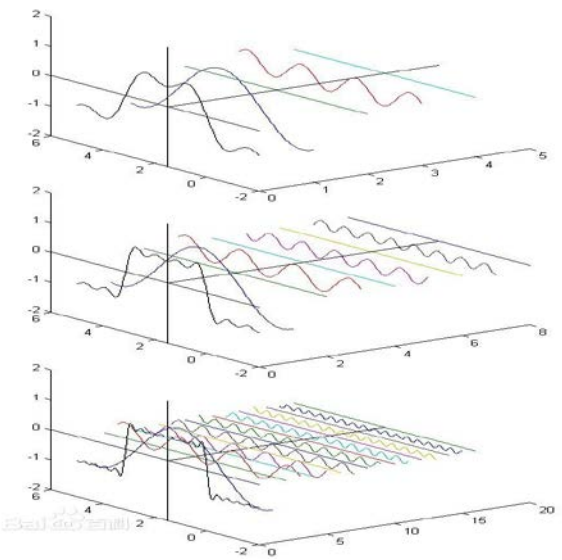


图 3-9　Matlab 的拟合结果

虽然图 3-10 中的心电信号波形从形态上看起来并不是正常的波形，但在 MIT BIH 的数据中有大量这种类型的数据，不能忽略不计。既然傅里叶级数对于这种数据不能很好地拟合，那么就不能采用该方法。如表 3-7 所示，我们固然可以通过增加阶数来提高拟合精度，但从表 3-8 中可以看出，6 阶傅里叶级数对于某些波形来说，拟合程度已经接近于 1，而且随着阶数的增加，参数的个数和拟合时间也会不断增加，如表 3-9 所示，经过笔者实验发现，使用一个 5 阶的傅里叶级数拟合一个波形所需时间约为 0.02～0.04 秒，按照 11 万个数据来算，拟合完整个数据集就需要 55 分钟左右，这对于整个实验流程显得过于漫长。

表 3-7　拟合参数表

拟合参数	3 阶	4 阶	5 阶	6 阶
ω	0.1554	0.1518	0.1413	0.1415
a_0	1112	1112	1111	1111
a_1	–17.93	–18.94	–21.67	–21.54
b_4	–22.54	–20.66	–14.74	–14.89
a_2	0.9052	5.518	17.98	17.76
b_2	27.76	26.96	20.38	20.64
a_3	11.96	8.246	–7.51	–7.145
b_3	–12.42	–17.13	–17.32	–17.58
a_4		–8.728	–2.745	–2.836
b_4		1.97	10.29	9.897
a_5			5.606	6.231
b_5			–2.011	–1.913

资料来源：马国伟. 基于聚类的心电信号分类方法研究.华东理工大学. 2012 年.

表 3-8　傅里叶级数阶数表

傅里叶级数阶数	SSE	RMSE	R-square
3 阶	3468	8.59	0.9227
4 阶	1435	5.647	0.968
5 阶	670.7	3.949	0.9851
6 阶	227.2	2.354	0.9949

资料来源：马国伟. 基于聚类的心电信号分类方法研究.华东理工大学. 2012 年.

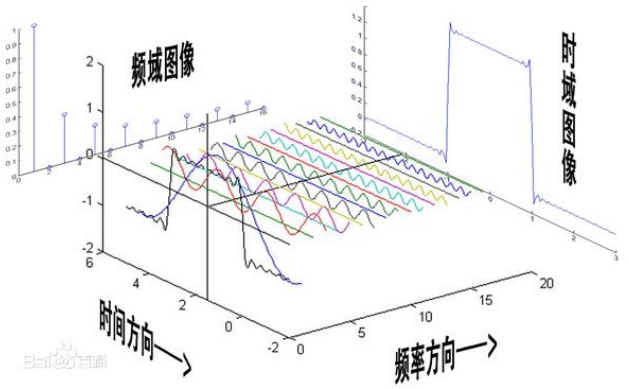


图 3-10　类直线波拟合结果

表 3-9　傅里叶级数拟合参数表

拟合参数	拟合结果	拟合参数	拟合结果
ω	0.01939	a_0	-1.356e+012
a_1	1.992e+012	b_1	1.388e+012
a_2	-6.02e+011	b_2	-1.628e+012
a_3	-2.453e+011	b_3	9.489e+011
a_4	3.245e+011	b_4	-2.8e+011
a_5	-1.395e+011	b_5	1.505e+010
a_6	2.852e+010	b_6	1.54e+010
a_7	-2.159e+009	b_7	-4.197e+009
a_8	-4.016e+007	b_8	3.24e+008

资料来源：马国伟. 8 阶傅里叶级数拟合参数.华东理工大学硕士论文. 2012 年.

我们把 55 个有效数据点作为一条记录去进行函数拟合，结果发现利用傅里叶级数拟合的方法对于大多数类似于正常的QRS波群能够很好地进行拟合，一个6阶的傅里叶级数拟合精度就能达到0.99以上，而且函数的参数只有 13 个，大大降低了数据的维数，但是当数据波形上下幅值波动很小，类似直线时，傅里叶级数的拟合效果就非常不好，即使是使用 8 阶的傅里叶级数，R-square 依然不能达到 0.9，且从拟合出来的波形外观上观察也和原始数据的波形有较大区别，拟合出来的参数绝对值过大。作为一个有效的特征提取方法，必须对所有类型的数据都能得到一个准确的结果，不然就不能够真实地还原出原始数据的特征，对数据分类的结果必然会造成很大影响。再者，本文中采用的数据量较大，特征提取的时间必须得到控制，不然这个实验周期就过长，从时间角度来看，拟合完这一整个数据集需近 1 个小时时间。如表 3-10 所示，在本文研究的内容中，不能采用傅里叶级数拟合这个方法来对数据进行特征值提取。

表 3-10　8 阶傅里叶级数拟合指标

拟合指标	指标结果
SSE	46.19
RMSE	1.17
R-square	0.8459

⑤ 基于原始数据的特征提取

傅里叶级数对于本文的研究内容来说并不是一个很好的选择，那么理所当然地会想到是否可以更换一个函数来用作拟合函数去提取数据的特征值，但经过对傅里叶级数方法的总结，拟合的时间问题必然会成为关键的环节。因此无论采用什么样的函数去拟合，无论结果有多准确，如果不能有效解决时间这个问题，依然不能称之为有效的特征提取方法。在没有很好的方法得到一个既要效果好，又要拟合时间短的函数的情况下，只能放弃使用拟合。

原始数据是分类的基础，它包含了心电信号最基本的信息，对特征值的提取也是从原始数据中取到对分类有用的数据，去掉原始数据中对分类无关或者影响不大的那部分。既然我们可以根据原始数据画出心电信号的波形图，那么我们自然也可以根据原始数据来直接分类，唯一的缺点就是数据的维度偏大或者数据中含有噪声。但是 55 维数据对于计算机来讲，计算压力不大，真正对计算有要求的是多达 11 万的海量数据。出于以下理由，本文最终采用了原始数据作为分类特征值。

⑥ 聚类算法的实现

聚类的准则函数一旦确定后就需要找出一种最优的划分，使得准则函数能够取到极小值。理论上我们可以采用穷举法，不断地尝试各种划分结果求出使准则函数取到极小值的分类，但穷举法对于大多数聚类问题是完全行不通的，因为假设有 100 个样本需要分成 5 类，那么就有约 106 种划分情况。所以通常情况下常用迭代最优化的方法来求出最优划分。其思想就是先找出一个初始划分或类别，再对其进行不断地调整聚类中心重新聚类，直到满足要求为止。这是一个动态的迭代过程，因此也称为动态聚类方法。K-Means 算法就是一个典型的动态聚类法。

K-Means 算法即 K-均值算法，是一种硬聚类算法，是典型的局域原型的目标函数聚类方法的代表，最早是由 James MacQueen 在 1967 年提出来的，是目前通过最小化某个准则函数进行聚类的方法中研究最多最广的一种聚类算法，许多学者通过基础的 K-Means 算法提出了很多新的改进的 K-Means 算法。典型的 K-Means 算法以平方误差准则实现了数据集的聚类，并且对于大数据集的处理效率也非常高。

K-Means 聚类问题的假设是有一组 N 个数据的集合 $X=\{x_1, x_2, x_3, \cdots, x_n\}$ 待聚类。K 均值值聚类问题是要找到 X 的一个划分 $P_k=\{C_1, C_2, C_3, \cdots, C_k\}$，使目标函数 $f(P_k)=\sum_{i=1}^{k}\sum_{x_i\in c_i} d(x_i, m_i)$ 最小。其中，$m_i=1/n_i$，$\sum_{x_i\in C_i} x_i$ 表示第 i 个簇中心位置，$i=1\cdots$，k；n_i 是簇 C_i 中数据项的个数；$d(x_i, m_i)$ 表示 x_i 到 m_i 的距离。通常的空间聚类算法是建立在各种距离基础上的，如欧几里得距离、曼哈顿距离和明考斯距离等。其中，最常用的是欧几里得距离。K-Means 算法的基本思想是：

给定一个包含 n 个数据对象的数据库，以及要生成簇的数目 k，随机选取 k 个对象作为初始的 k 个聚类中心；然后计算剩余各个样本到每一个聚类中心的距离，把该样本归到离它最近的那个聚类中心所在的类，对调整后的新类使用平均值的方法计算新的聚类中心；如果相邻两次的聚类中心没有任何变化，说明样本调整结束且聚类平均误差准则函数已经收敛。本算法在每次迭代中都要考察每个样本的分类是否正确，若不正确，就要调整。在全部样本调整完成后修改聚类中心，进入下一次迭代。如果在一次迭代算法中，所有的样本被正确分类，则不会有调整，聚类中心不会有变化。在算法迭代中值在不断减小，最终收敛至一个固定的值。该准则也是衡量算法是否正确的依据之一。

Matlab 函数的核心代码如下：

```
[IDX, C, SUMD, D]=kmeans (X, K, 'PARAM 1', valI, 'PARAM2', va12,…)
```

IDX 是样本的类别号，C 表示 K 个质心点的位置，SUMD 是一个 $1\times K$ 的矩阵，表示了类间所有的点与该类质心点距离之和，D 是一个 $N\times K$ 的矩阵，N 表示样本个数，D 表示了每个样本与所有质心的距离。X 是 $M\times N$ 的矩阵，由 M 个样本组成，每个样本有 N 个字段的数据集。参数 PARAM 1 的值为 Distance（距离测度），Distance 对应的值有：sqEuclidean，欧氏距离(默认时，采用此距离方式)；cityblock，绝度误差和等。参数 PARAM2 的值为 Start(初始质心位置选择方法)，其对应的值有：sample，从 X 中随机选取 K 个质心点；`uniform'，根据 X 的分布范围均匀地随机生成 K 个质心；cluster，初始聚类阶段随机选择 10%的 X 的子样本(初始使用 sample 的方法)；matrix 提供一个 $K\times P$ 的矩阵，作为初始质心位置集合。此外，kmeans 函数的参数还可以指定聚类重复次数，每次聚类的最大迭代次数等。

本节中选取欧氏距离作为样本间距离计算方法，初始聚类中心采用随机 z 生成中心点，将 112601×55 的样本数据矩阵进行 K-Means 算法聚类，将样本数据分成 30 类，为了使结果尽可能收敛，但也不耗费过长时间，本文设定迭代次数为 1000 次，实际可能并不需要这么多，重复实验 5 次，取结果中 J 最小的一次作为实验结果，整个实验过程耗时约 6 分钟（K-Means 算法的聚类结果和实验数据如表 3-11 和表 3-12 所示）。分类效果如图 3-11 所示。

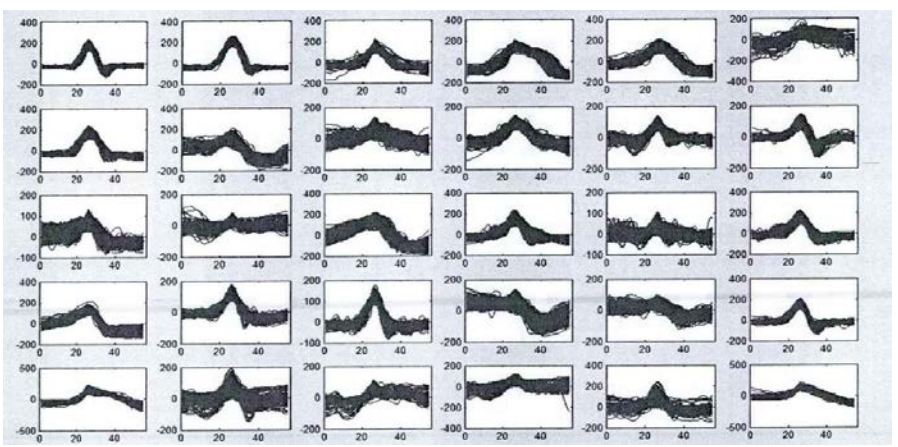


图 3-11　K-Means 算法的聚类结果（马国伟，华东理工大学，2012 硕士论文）

表 3-11　K-Means 算法的聚类结果

类别号	样本数	聚合度	类别号	样本数	聚合度	类别号	样本数	聚合度
1	5188	111	2	1826	209	3	3619	164
4	946	769	5	2982	175	6	1245	494
7	4280	196	8	3131	347	9	1912	302
10	3866	110	11	6825	90	12	6842	114
13	9273	103	14	3738	129	15	1172	854
16	3037	212	17	6049	120	18	5162	238
19	2613	277	20	3459	121	21	7984	87
22	2690	328	23	4618	157	24	6536	156
25	1696	295	26	4440	196	27	2356	80
28	3102	196	29	220	1227	30	1794	189

资料来源：马国伟. K-Means 算法聚类结果数据. 华东理工大学. 2012 硕士论文.

表 3-12　K-Means 算法实验数据

样本个数	112601
类别数	30
迭代次数	1000
耗时	6 分钟
聚类准则	8046
平均类内聚类准则	268.2

资料来源：马国伟. K-Means 算法实验数据. 华东理工大学. 2012 硕士论文.

3.4.3　案例

本案例中有很多值得关注的数据处理、技术逻辑、算法运筹、模式识别的亮点，我们不妨来看一下。

① 在对波形数据的预处理过程中，作者采用了大胆的解构主义方法创造性地截取 P 段、QRS 段，以峰值点 R 点为中心左右共取 55 个点（参见图 3-6 和图 3-8），这样的数据预处理方法不但有效降维数据集，节省了计算与缓存资源，也是电生理信号处理的娴熟手段。由于 P 波、QRS 波更具有临床意义，这样的降维截取有一定的可取之处，在处理 11 万份电生理信号大数据的条件下，这样的截取与重构是必要的。在计算机化的实操过程中，一个个电生理脉冲波形图可以被还原成一个个数据流，这就实现了图形的数据化处理。数据的预处理也充分体现了数据挖掘的本质就是计算机化的数据机器处理，这和传统的统计学分析有很大的区别，因此认为数据挖掘不过是传统统计学的变种的观点显然是站不住脚的。数据挖掘与统计学有很多的交集，但本质上两者有很大的不同。当然，

完全脱离统计学的数据挖掘技术也不存在。

② 傅里叶变换无疑是法国数学家对脉冲信号处理最有力的武器。傅里叶变换的精妙之处在于把一个个波形图分解为正弦波与余弦波的正交，也就是把波函数转换为三角函数，把一个脉冲信号问题变成可以拟合的数学问题，其本质还是积分问题，一个多子集的求和问题。案例中的作者试图用傅里叶变换的经典方法去拟合波形图，遗憾的是拟合的结果并不成功（见图 3-10）。案例作者重点强调了数据维数与运行时间的约束条件导致傅里叶拟合的失败。事实上拟合不好的原因很多，比如噪声剔除不够、函数需要修正、维数需要限制等因素，结果是选用了原始的图像数据作为特征值处理。这样的数据特征处理在实践中是经常出现的，很有效果但在数据特征值的提取上有不严谨的地方。从一堆数据集中提取有代表性的特征值是聚类中最重要的步骤之一，各个数族中心点的距离远近，其本质是按特征、按维数来自组织聚类的。我们同时也通过本案例的拟合方法看到了数学中微积分思想的伟大，无论是微分过程还是积分过程，数据的拟合思想方法把圆与弧度的问题变为直线与点的求和。

③ 如何理解 K-Means 算法的基本原理是很多初学者感到很难的问题。实际上 K-Means 这样一种均值的聚类算法最主要的思想就是，首先假设有若干个数据队列集合等待聚类，随机选择 *K* 个对象为第一次聚类的中心点，依据每个元素相对于中心点 K 的距离远近进行聚类，然后按照调整后的新类使用平均值的方法计算新的聚类中心；如果相邻两次的聚类中心没有任何变化，说明样本调整结束且聚类平均误差准则函数已经收敛。这样多次迭代后就可以确保每一个样本的正确聚类。

④ 本案例中的数据选择也是别具特色的，采用 MIT-RIH 麻省理工学院的心律失常国际数据库调取 11 万份心电数据有效破解了国内心电大数据共享困难的难题。最终的聚类结果（见图 3-11）产生了心电数据的 30 个分型，这也是模式识别中解释性建模、描述性建模的最生动的诠释。从解释性建模角度看，心电数据的正常与异常模式的无限逼近与拟合成为机器学习的重要一步，追求精确性建模目标是模糊建模理论近年来追求的主要目标，在心电大数据中如何精准地识别每一份心电数据的正常或异常就成为心电数据挖掘是否成功的试金石。从描述性建模看，心电数据的分型与模块化聚类有效解决了从特殊到一般的规律性总结，机器学习的最大障碍被破除，人工智能判读心电数据最重要的模式识别方法：心电大数据的基本分型已经完成。当然，我们对于机器学习与人工智能的漫漫征途也不能太乐观，实际上心电波形图的判读是十分复杂的人脑活动，有极强的专业性与经验性，即使是 ST 波不显著波形的判读对于有经验的医生来讲也能抓住蛛丝马迹来判断心电的异常情况。心电数据的判读从来都不是大众可以完成的事情。毕竟，11 万份心电大数据的 K-Means 聚类让我们看到了曙光。

3.5 贝叶斯算法的应用案例

3.5.1 一个流传甚广的故事

现分别有 A、B 两个容器，在容器 A 里分别有 7 个红球和 3 个白球，在容器 B 里有 1 个红球和

9 个白球，现已知从这两个容器里任意抽出了一个球，且是红球，问这个红球是来自容器 A 的概率是多少?

假设已经抽出红球为事件 B，从容器 A 里抽出球为事件 A，则有：$P(B) = 8/20$，$P(A) = 1/2$，$P(B|A) = 7/10$，按照公式，则有：$P(A|B) = (7/10)\times(1/2) / (8/20) = 0.875$。

贝叶斯开启了不确定性问题的解决方案，成为统计学历史上的飞跃，也终结了统计学大多数解决确定性问题的历史，开启了概率论的新篇章。概率工具的出现拓展了人类对数据世界的认识视野，在天气预报、医学疾病预测等科学研究中大有用武之地。概率使用大数定理或小概率事件的描述方法改善了人们对某些物理或生命现象的描述方法，成为解释性建模、描述性建模的又一革新理念与工具。

毫不夸张地说，贝叶斯理论的出现对疾病数据的研究如虎添翼。

贝叶斯公式为利用搜集到的信息对原有判断进行修正提供了有效手段。在采样之前，经济主体对各种假设有一个判断（先验概率），关于先验概率的分布，通常可根据经济主体的经验判断确定（当无任何信息时，一般假设各先验概率相同），较复杂精确的则可利用包括最大熵技术或边际分布密度以及相互信息原理等方法来确定先验概率分布。

3.5.2 一个贝叶斯算法的医学案例

1. 本案研究目标的 7 个主要方面

① 设计了用于高血压诊断的本体模型，采用本体模型和贝叶斯网络模型之间的映射关系来自动生成贝叶斯网络模型，分析了贝叶斯网络模型输入变量的量化方式以及缺失值处理方式。

② 阅读领域文献并与内科医生交流以收集高血压患者心血管风险水平分类的知识，然后将这些知识采用 SWRL 语言进行描述以建立用于分类任务的知识库。

③ 分析为了达到心血管风险水平分类目的需要的体检数据以建立高血压电子病历数据库。

④ 采用自顶向下、逐步细分的方式，利用斯坦福大学开发的 Protege 工具构建了用于心血管风险水平分类的本体。

⑤ 研究利用数据库记录来自动生成本体实例的方式以充分利用数据库和本体的优势。

⑥ 在分析本体理论与规则推理的基础上，提出高血压患者心血管风险水平分类系统框架，并利用 Pellet 推理引擎实现系统的诊断功能。

⑦ 采用 Java 语言和 SSH 框架将前面的研究整合成一个界面友好的诊断支持系统。

2. 研究方法与理论

① 本体论

Perez 在他的论文中总结出本体包含 5 个基本的建模原语。这些原语是：类、关系、函数、公理、实例。我们通常也把类说成概念。概念有非常广泛的含义，可以指任何事物，例如高血压、血常规、血生化、个人史等。关系代表了概念在领域之中的相互联系，例如子类关系、逆关系。关系在形式上可定义成 n 维笛卡儿积的子集。函数也是一类关系，不过它比较特殊，类似数学函数，由一部分内容能推导出另一部分的结果，例如有两个变量 x 和 y，x 可以通过函数 Son-of(x,y)唯一确定它的父

亲 y。公理是那些无须去证明的客观事实或规律，代表永真断言，例如人是有寿命的。实例是类的对象，代表元素，有些类似面向对象的程序语言中的对象。

② 贝叶斯网络模型

本文研究的目标是能够对高血压患者的心血管风险水平进行分类（如表 3-13～表 3-15 所示）。而分类的前提是已经确诊病人患有高血压，对非高血压患者这个分类是毫无意义的。因此，在对患者的心血管风险水平进行分类之前，我们需要诊断该患者是否是高血压患者。诊断操作可定义成这样一个过程：鉴别建模问题域的一系列假设，并从这些假设中找出能以最高概率匹配现实世界状况的那个假设。在医学诊断中，不确定性来自于信息的不完整或者不可靠，或者知识的不一致性导致决策者不能评估假设的真实程度。

本文选用本体和贝叶斯网络的方法来处理知识管理以及不确定性。本体能够将领域知识表达成机器可读的形式。它能够表达大型、复杂的领域的组织结构，但是它不能够处理不确定性，这是本体应用的一个短板。贝叶斯网络在解决不确定性知识的置信度时非常有效，适用于不确定知识的表示及推理。为了克服彼此的缺点，本体和贝叶斯网络能够互补，因此，可以创建一个本体驱动的贝叶斯网络模型。此外，本体的来源非常广泛，尤其是自动化本体构建技术出现后，例如从关系数据库自动生成本体可以使本体的构建自动化。我们将本体驱动的贝叶斯网络模型作用于自动构建的本体上，也可以实现贝叶斯网络构建的自动化。

理论依据：本体拥有自己的实例，这些实例通过对象关系属性形成一个图结构。

贝叶斯网络拥有自己的变量，这些变量之间通过依赖关系也能形成一个图结构。因此，可以将本体实例映射成贝叶斯网络变量，将本体实例之间的对象关系映射成贝叶斯网络变量之间的依赖关系。故根据本体文件来自动构建贝叶斯网络是可行的。

因果独立性：在贝叶斯网络模型中，若父变量 $X1$、……、$X2$，X_n 均可以单独对子变量 C 产生概率影响，则称这 n 个父节点是相互因果独立的。在因果独立的情况下，贝叶斯网络模型称为 Noisy-or 模型，Noisy-or 模型中各节点的父节点之间都是因果独立的。

表 3-13　其他危险因素对高血压患病率的影响

危险因素	危险因素水平	患病率（%）
酒精摄入量（g/d）	<4.8	24.04
	>=4.8 和<10.51	23.65
	>=10.51 和<19.94	26.25
	>=19.94 和 40.03	30.20
	>=40.03	35.22
体重指数（kg/m^2）	<=18.5	13.70
	>18.5 和<=25	16.50
	>25 和<=28	33.30
	>28	51.20

续表

危险因素	危险因素水平	患病率（%）
高血压家族史	无	18.22
	有	30.38
甘油三酯	正常	20.69
	偏高	37.2
高密度脂蛋白	正常	22.68
	偏低	25.47
胆固醇	正常	21.29
	偏高	43.26
吸烟	否	22.54
	是	26.32
高血糖	否	22.35
	是	64.39

资料来源：黄飞. 基于贝叶斯网络和本体的高血压患者心血管风险水平分类系统研究.太原理工大学. 2014 年硕士论文.

表 3-14　体重指数与高胆固醇、高甘油三酯、高血糖患病率的关系

体重指数（kg/m^2）	胆固醇偏高（%）	甘油三酯偏高（%）	高血糖（%）
BMI<=18.5	17.65	5.88	0.00
BMI<=25	20.81	27.27	3.59
BMI<=28	27.85	49.12	6.14
BMI>28	33.33	62.82	14.74

表 3-15　性别对吸烟、喝酒概率的影响

		酒精摄入量（g/d）				
性别	吸烟（%）	<4.8	<10.51	<19.94	<40.03	>=40.03
男	59.99	13.3	4.3	26.5	25.5	30.4
女	3.1	32.4	2.1	28.0	27.0	10.5

3. 实验过程

本文采用 Netica 来辅助诊断操作。Netica 既提供了可以操纵贝叶斯网络的多种程序语言 API，又提供了可视化操作贝叶斯网络的软件产品。实验方法如下。

（1）生成贝叶斯网络层次结构

建立了用于高血压诊断的领域本体，该本体存储在一个符合 OWL 规范的 XML 文件中。该文件将成为本章讲解的从疾病本体自动构建疾病的贝叶斯网络模型算法的输入，同时该算法还将输出一个 Netica 软件产品可识别的贝叶斯网络模型。自动生成的贝叶斯网络模型如图 3-12 所示。

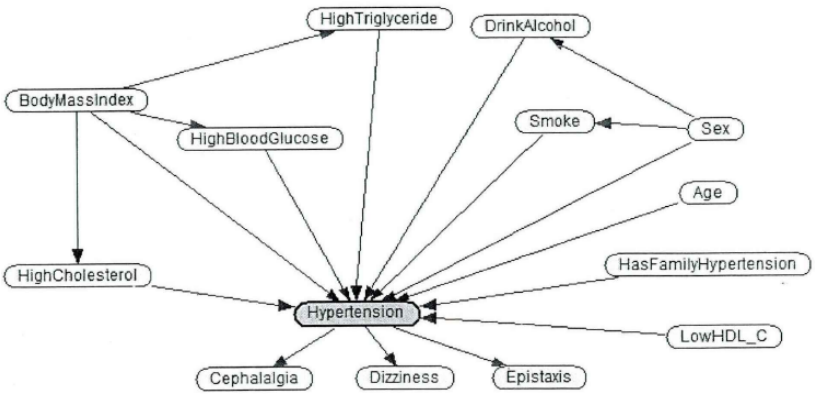


图 3-12　自动生成的贝叶斯网络模型

（2）输入变量

本文的贝叶斯网络有 10 个输入变量，现在我们来确定每个输入变量的成员函数，成员函数决定了每个对象在贝叶斯网络中对应的成员。

① 体重指数节点（BodyMassIndex）

体重指数被分成了 4 个区域，分别对应的是消瘦、正常、超重、肥胖，它们对应的体重指数(单位：kg/m^2)区域如下：

消瘦=BMI<=18.5

正常=18.5< BMI<=25

超重=25< BMI<=28

肥胖=BMI>28

用户在输入体重指数节点证据时只需在贝叶斯网络中选择相应的区域即可。

② 胆固醇偏高节点（HighCholesterol）

胆固醇偏高节点是一个布尔型节点，在贝叶斯网络中只需选择 yes 和 no 两个选项。不过我们可以对胆固醇偏高的概率进行度量（即对 yes 的概率进行度量）。我国标准建议胆固醇小于或等于 5.2mmol/L 为正常，大于 5.2mmol/L 并且小于等于 5.66mmol/L 为临界升高，大于 5.66mmol/L 为升高。

③ 血糖偏高节点是一个布尔型节点，在贝叶斯网络中只需选择 yes 和 no 两个选项。我国标准建议空腹血糖值 3.89～6.1 mmol/L 为正常，大于 6.1 mmol/L 并且小于 7.0mmol/L 为空腹血糖受损，大于等于 7.0mmol/L 考虑糖尿病。

④ 甘油三酯偏高节点是一个布尔型节点，在贝叶斯网络中只需选择 yes 和 no 两个选项。我国标准建议甘油三酯小于 1.7mmol/L 为正常，大于等于 1.7mmol/L 并且小于 5.65mmol/L 为临界升高，大于等于 5.65mmol/L 为明确升高。

⑤ 高密度脂蛋白偏低节点是一个布尔型节点，在贝叶斯网络中只需选择 yes 和 no 两个选项。我国标准建议高密度脂蛋白胆固醇大于 1.16mmol/L 为正常。小于等于 1.16mmol/L 并且大于等于 0.93mmol/L 为临界水平，小于 0.93mmol/L 为明确降低。

⑥ 饮酒节点（DririkAlcohol）

病人每日的酒精摄入量（单位：g/d）被划分成 5 个区域，对应的酒精摄入量如下。

区域 1= DrinkAlcohol<4.8

区域 2= 4.8=<DrinkAlcohol<10.51

区域 3= 10.51=<DrinkAlcohol<19.94

区域 4= 19.94=<DrinkAlcohol<40.03

区域 5= DririkAlcohol>=40.03

用户在输入病人每日的酒精摄入量这一证据时只需在贝叶斯网络中选择相应的区域即可。

⑦ 有高血压家族史节点（HasFamilyHypertension）

有高血压家族史节点是一个布尔型节点，在贝叶斯网络中只需选择 yes 和 no 两个选项。该节点无须对 yes 的概率进行度量，只要病人的父母有任何一个是高血压患者，那么该节点为 yes 的概率就是 1.0，否则是 0.0。

⑧ 吸烟节点（Smoke）

吸烟节点是一个布尔型节点，在贝叶斯网络中只需选择 yes 和 no 两个选项。该节点无须对 yes 的概率进行度量，只要病人目前抽烟而不管量的大小，那么该节点为 yes 的概率就是 1.0，否则是 0.0。

⑨ 性别节点（Sex）

性别节点有两个状态分别是 male 与 female。

⑩ 年龄节点（Age）

年龄节点被划分成 6 个区域，对应的年龄如下.

区域 1=Age<25

区域 2=25=<Age<3 5

区域 3=35=<Age<45

区域 4=45=<Age<55

区域 5=55=<Age<65

区域 6=Age 大于等于 65

用户在输入病人年龄时只需在贝叶斯网络中选择相应的区域即可（表 3-16）。

表 3-16　高血压诊断准确性对比表

类　　别	临床确诊	贝叶斯网络确诊
高血压患者	52	60
正常人	65	57
合计	117	117

资料来源：黄飞. 基于贝叶斯网络和本体的高血压患者心血管风险水平分类系统研究.太原理工大学. 2014 年硕士论文.

敏感性与特异性是一对矛盾体，提高其中一个的精度必然降低另外一个的精度。

本文的敏感性和特异性精度主要与 3 个因素有关：确诊高血压的阈值(本文设置的阈值为 80%)、条件概率表的准确性、贝叶斯网络结构的合理性。

此外，我们还可以在确定病人患高血压的概率后，进一步推理出该病人由高血压引起的症状或者病症的概率。本文尚未提供高血压与这些节点之间关系的条件概率表。

4. 小结

本章提出的本体驱动的贝叶斯网络模型能够自动解析本体中实例之间的层次关系，来动态地构建贝叶斯网络，充分利用了本体表达能力强并且具有语义性的优点。另外，该模型也利用了贝叶斯网络能够解决不确定性和非完整性信息的优势。

向本体文件中添加新的实例后无须做任何代码更改，只需重新读取本体文件即可更新贝叶斯网络模型。本文提出的医学本体模型（如图 3-12 所示）具有一定的通用性，可以很容易地进行本体扩充并将其用于其他疾病的诊断。

本章采用经过量化的高血压门诊数据来验证该模型的有效性。实验结果表明该模型能够根据我们所提供的病人证据做出正确诊断。由于现有的依靠测量诊室血压来诊断高血压的方法存在一些缺陷，因此，本章建立的高血压诊断模型可以作为现有诊断高血压方法的补充，辅助内科医生更好地做出诊断。

5. 数据库表单设计（表 3-17 ~ 表 3-20）

表 3-17　病人基本信息表

序　号	列　名	中文名称	类　型	长　度	小　数	允许为空
1	Id	病人 ID	Char	36	0	0
2	Register Date	登记日期	Date	7	0	1
3	Name	姓名	Varchar2	20	0	0
4	Sex	性别	Char	2	0	0
5	Birthday	出生日期	Date	7	0	0
6	Address	住址	Varchar2	50	0	1
7	Telphone	联系电话	Char	20	0	1
8	Email	电子邮件	Varchar2	40	0	1
9	Username	用户名	Varchar2	20	0	0
10	Password	密码	Varchar2	32	0	0
11	Visible	是否可见	Number	1	0	1

表 3-18　医务人员信息表

序　号	列　名	中文名称	类　型	长　度	小　数	允许为空
1	Id	病人 ID	Char	36	0	0
2	CertificateId	医师资格证编号	Char	27	0	0
3	Name	姓名	Varchar2	20	0	0
4	Sex	性别	Char	2	0	0
5	RgisterDate	入职日期	Date	7	0	0
6	Birthday	出生日期	Date	7	0	0
7	Address	住址	Varchar2	50	0	0
8	Telphone	联系电话	Char	20	0	0
9	Email	电子邮件	Varchar2	40	0	0
10	Username	用户名	Varchar2	20	0	0
11	Password	密码	Varchar2	32	0	0
12	Department	所属科室 ID	Varchar2	30	0	0
13	Iden_card	身份证编号	Char	18	0	0
14	Visible	是否可见	Number	1	0	1

表 3-19　血压测量记录

序　号	列　名	中文名称	类　型	长　度	小　数	允许为空
1	Id	记录 ID	Char	36	0	0
2	Patient_ id	病人 ID	Char	36	0	0
3	Checkup_date	测量日期	Date	7	0	0
4	Sys_blood_pre	收缩压	Number	3	0	0
5	Dia_blood_pre	舒张压	Number	3	0	0
6	Blood_category	血压分类	Varchar2	20	0	1

表 3-20　血生化指标

序　号	列　名	中文名称	类　型	长　度	小　数	允许为空
1	patient_id	病人 ID	Char	36	0	0
2	checkup_date	采集日期	Date	7	0	0
3	potassium	钾	Number	3	1	1
4	sodium	钠	Number	3	0	1
5	uric_acid	尿酸	Number	3	0	1
6	fas_blood_glu	空腹血糖	Number	3	1	1
7	two_hour_glu	餐后 2h 血糖	Number	3	1	1
8	serum_creatinine	血肌酐	Number	3	0	1

续表

序　号	列　名	中文名称	类　型	长　度	小　数	允许为空
9	urea_nitrogen	尿素氮	Number	3	1	1
10	total_cholesterol	总胆固醇	Number	3	1	1
11	triglyceride	甘油酯	Number	4	2	1
12	hdi_cholesterol	高密度脂蛋白胆固醇	Number	4	2	1
13	ldl_cholesterol	低密度脂蛋白胆固醇	Number	4	2	1
14	glutamic_pyr_tra	谷丙转氨酶	Number	3	0	1
15	doctor_id	采集医生编号	Char	36	0	1

资料来源：黄飞. 基于贝叶斯网络和本体的高血压患者心血管风险水平分类系统研究.太原理工大学. 2014 年硕士论文.

数据库表除了有相互关联之外，每个表都有自己的字段信息，这些字段信息约束着保存在该表的记录的属性，如控制数据类型、约束数据长度、是否允许为空等。表格中允许为空这列只有 0 和 1 两种状态，0 表示不允许为空，1 表示可以为空。

6. 本体知识库的构建（表 3-21 ~ 表 3-27）

表 3-21　静态数据实例表

类　名	实　例
靶器官损害	颈动脉超声 IMT>0.9mm 或动脉粥样斑块、左心室肥厚、劲-股动脉脉搏波速>12m/s、估算的肾小球滤过率降低或血清肌酐轻度升高、踝/臂血压指数<0.9、微量白蛋白尿
心血管危险因素	高血压（1 ~ 3 级）、高龄、糖耐量受损和（或）空腹血糖异常、吸烟、血脂异常、腹型肥胖或肥胖、早发心血管病家族史、血同型半胱氨酸升高
临床疾患	脑血管病、肾脏疾病、心脏疾病、视网膜病变、外周血管疾病、糖尿病

表 3-22　血压水平分类与定义

分　类	舒张压（mmHg）		收缩压（mmHg）
正常血压	<80	和	<120
正常高值血压	>=80 和<90	和（或）	>=120 and<140
高血压	>=90	和（或）	>=140
1 级高血压	>=90 和<100	和（或）	>=140 and<160
2 级高血压	>=100 和<110	和（或）	>=160 and<180
3 级高血压	>=110	和（或）	>=180
单纯收缩期高血压	<90	和	>=140

表 3-23　血压水平分类知识示例

示例 1：	
知识表示	BloodPressure(?b)^hasDiaBloodPre(?b,?d)^hasSysBloodPre(?b,?s)^lessThanOrEqual(?d,89) ^greaterTanOrEqual(?s,120)^ lessThanOr Equal(?s,139)→hasBloodCategory(?b,“正常高值血压”)
知识解释	如果：b 是一个血压测量记录；d 是这个血压测量记录中的舒张压，并且舒张压 d 小于或等于 89；s 是这个血压测量记录中的收缩压，并且收缩压 s 大于或等于 120，小于或等于 139 则：血压测量记录 b 可以分类成“正常高值血压”
示例 2：	
知识表示	BloodPressure(?b)^hasDiaBloodPre(?b,?d)^hasSysBloodPre(?b,?s) greaterTanOrEqual(?d,89)^ lessThanOr Equal(?s,139)→hasBloodCategory(?b,“正常高值血压”)
知识解释	如果：b 是一个血压测量记录；d 是这个血压测量记录中的舒张压，并且舒张压 d 大于或等于 80，小于或等于 89；s 是这个血压测量记录中的收缩压，并且收缩压 s 小于或等于 139 则：血压测量记录 b 可以分类成“正常高值血压”
示例 3：	
知识表示	BloodPressure(?b)^hasDiaBloodPre(?b,?d)^hasSysBloodPre(?b,?s)^greaterThanOrEqual(?d,90)^lessThanOrEqual (?d,99) ^ lessThanOr Equal(?s,159)→hasBloodCategory(?b,“1 级高血压（轻度）”)
知识解释	如果：b 是一个血压测量记录；d 是这个血压测量记录中的舒张压，并且舒张压 d 大于或等于 90，小于或等于 99；s 是这个血压测量记录中的收缩压，并且收缩压 s 大于或等于 159。 则：血压测量记录 b 可以分类成“1 级高血压（轻度）”

表 3-24　心血管预后重要因素知识库

类　别	判断依据
心血管危险因素	男性>55 岁，女性>65 岁 高血压（1～3 级） 吸烟 血脂异常：TC≥5.7mmol/L（220mg/dL）或 LDL-C>3.3 mmol/L（130mg/dL）或 HDL-C<1.0 mmol/L（40mg/dL） 糖耐量受损（2 小时血糖 7.8～11.0 mmol/L）和/或空腹血糖异常（6.1～6.9 mmol/L） 腹型肥胖（腰围：男性≥90cm，女性≥85cm）或肥胖（BMI≥28kg/m^2） 早发心血管病家族史（一级亲属发病年龄<50 岁） 高同型半胱氨酸>10 mmol/L
靶器官损害	劲动脉超声 IMT>0.9mm 或动脉粥样斑块 左心室肥厚：心电图（Sokolow-Lyons>38mv 或 Cornell>2440mm・mms）；超生心动图 LVMI（男>=125，女>120g/m^2） 颈 9 股动脉脉搏波速度>12m/s(*选择使用) 估算的肾小球搏波过率降低(eGFR<60ml/min/1.73m^2)或血清肌酐轻度高（男性 115～133umo1/L，女性 107～124unol/L） 踝/臂血压指数<0.9（*选择使用）

续表

类　　别	判断依据
	微量白蛋白尿（30～300mg/24h）或白蛋白/肌酐比（≥30mg/g）
临床疾患	脑血管病（缺血性脑卒中、脑出血、短暂性脑缺血发作） 肾脏疾病（肾功能受损、糖尿病肾病、血肌酐：男性>133umol/L，女性>124 umol/L、蛋白尿>300 mg/24h） 心脏疾病（心绞痛、心肌梗塞史、充血性心力衰竭、冠状动脉血运重建史） 视网膜病变：出血或渗出、视乳头水肿 外周血管疾病 糖尿病（空腹血糖：≥7.0 mmol/L、餐后血糖≥11.1 mmol/L、糖化血红蛋白：≥6.5%）

表 3-25　逻辑判断知识库示例

示例 1：	
知识表示	PatientOnt(?p)^has(?p,?h)^HypertensionEMR(?h)^contains(?h,?l)^LaboratoryExam(?l)^contains(?l,?b)^Blood Biochemistry(?b)^hasTwo HourGlu(?b,?sc)^greaterThenOrEqual(?sc,124)^lessThanOrEqual(?thg, 11.0) → diagnosedWith(? P, Plood GlucoseProblem)
知识解释	如果：p 是一个病人；p 有高血压电子病历 h；h 含有实验室检查项目 l；l 含有血生化实例 b；b 的血清肌酸酐 sc 大于或等于 124（umol/L），病人 p 的性别是女 则：诊断该病人 p 患有肾病
示例 2：	
知识表示	PatientOnt(?p)^has(?p,?h)^HypertensionEMR(?h)^contains(?h,?l)^LaboratoryExam(?l)^contains(?l,?b)^Blood Biochemistry(?b)^hasSerumCreatinine(?b,?thg)^greaterThenOrEqual(?thg,7.8)^hasSex(?p,?sex)^stringEqualIgnore Case(?sex,“女”)→diagnosedWith(?p,KidneyDisease)
知识解释	如果：p 是一个病人；p 有高血压电子病历 h；h 含有实验室检查项目 l；l 含有血生化实例 b；b 餐后两小时血糖 thg 大于或等于 7.8（mmol/L）并且小于或等于 11.0（mmol/L） 则：诊断该病人 p 患有糖耐量受损
示例 3：	
知识表示	PatientOnt(?p)^has(?p,?h)^HypertensionEMR(?h)^GeneraExam(?g)^contatins(?g, ? ph)^PhysicalExam(?pg)^has Waistline(?ph, ?w)^greaterThanOrEqual(?w, 90)^hasSex(?p, ?sex)^stringEqualIgnoreCase(?sex, " 男 ") → diagnosedWith(?p, AbdominalObesity)
知识解释	如果：p 是一个病人；p 有高血压电子病历 h；h 含有一般检查项目 g；g 含有体格检查实例 ph；ph 有腰围 w；w 大于或等于 90（cm）病人 p 的性别是男 则：诊断该病人 p 患有腹型肥胖

表 3-26　高血压患者心血管风险判断标准

危险因素，靶器官损害及并发症个数	高血压级别		
	1 级	2 级	3 级
无	低位	中危	高危

续表

1～2 个其他危险因素	中危	中危	很高危
>=3 个其他危险因素或>=1 个靶器官损害	高危	高危	很高危
>=1 个临床并发症或患有糖尿病	很高危	很高危	很高危

表 3-27　风险判断示例

示例 1：	
知识表示	PatientOnt(?p)^has(?p,?hdr)^HypertensionDiagnosticRepor(?hdr)^hasCVDhazardNum(?hdr,?d)^ lessThanOrEqual(?d,2)^ hasTargetOrganDamageNum(?hdr,?e)^equal(?e,0)^ hasDiseaseNum(?hdr,?f)^ equal (?f,0)^has(?p,?emr)^HypertensionEMR(?emr)^contains(?amr,?ge)^GeneralExam(?ge)^contains(?ge,?bp)^ Blood Pressure(?bp,)^hasBloodCategory(?bp, ?bc)^stingEqualIgnoreCase(?bc, "2 级高血压（中度）")→ hasCVDHazardLevel(?hdr,“中危”)
知识解释	如果：p 是一个病人；p 有高血压诊断报告 hdr；hdr 中的心血管危险因素的数量 d 小于或等于 2；hdr 中的靶器官损害数量 e 等于 0；hdr 中的临床并发症数量 f 等于 0；p 有高血压电子病历 emr；emr 含有一般检查项目 ge；ge 含有血压测量记录 bp；bp 的血压水平归类 bc 为“2 级高血压（中度）” 则：有高血压诊断报告中将病人的心血管风险水平分类成“中危”
示例 2：	
知识表示	PatientOnt(?p)^has(?p,?hdr)^HypertensionDiagnosticRepor(?hdr)^hasCVDhazardNum(?hdr,?d)^ equal(?d,0)^ hasTargetOrganDamageNum(?hdr,?e)^equal(?e,0)^ hasDiseaseNum(?hdr,?f)^ equal (?f,0)^has(?p,?emr)^HypertensionEMR(?emr)^contains(?amr,?ge)^GeneralExam(?ge)^contains(?ge,?bp)^ Blood Pressure(?bp)^ hasBloodCategory(?bp,?bc)^stingEqualIgnoreCase(?bc, "3 级高血压（重度）")→ hasCVDHazardLevel(?hdr, "高危")
知识解释	如果：p 是一个病人；p 有高血压诊断报告 hdr；hdr 中的心血管危险因素的数量 d 等于 0；hdr 中的靶器官损害数量 e 等于 0；hdr 中的临床并发症数量 f 等于 0；p 有高血压电子病历 emr；emr 含有一般检查项目 ge；ge 含有血压测量记录 bp；bp 的血压水平归类 bc 为“3 级高血压（重度）” 则：有高血压诊断报告中将病人的心血管风险水平分类成“高危”

小知识：知识库

知识库（Knowledge Base）是知识工程中结构化、易操作、易利用、全面有组织的知识集群，是针对某一（或某些）领域问题求解的需要，采用某种（或若干）知识表示方式在计算机存储器中存储、组织、管理和使用的互相联系的知识片集合。这些知识片包括与领域相关的理论知识、事实数据，由专家经验得到的启发式知识，如某领域内有关的定义、定理和运算法则以及常识性知识等。

知识是人类智慧的结晶。知识库使基于知识的系统（或专家系统）具有智能性。并不是所有具有智能的程序都拥有知识库，只有基于知识的系统才拥有知识库。现在许多应用程序都利用知识，其中有的还达到了很高的水平，但是，这些应用程序可能并不是基于知识的系统，它们也不拥有知识库。一般的应用程序与基于知识的系统之间的区别在于：一般的应用程序是把问题求解的知识隐含地编码在程序中，而基于知识的系统则将应用领域的问题求解知识显式地表达，并单独地组成一

个相对独立的程序实体。

贝叶斯算法的出现无疑是革命性的，它的知名度不仅仅因为那个众所周知的故事，身为教会牧师的贝叶斯是一位业余的数学家，在牛顿万有引力的证明过程中，贝叶斯力挺牛顿，提供了很多的数学推导。令人唏嘘不已的是那个时代的数学家们早已经证明了“正向概率”事件，比如已知 A、B 两个口袋中 A 有黑球 8 个、红球 2 个，B 有红球 6 个、黑球 4 个，随机取出 4 个球，红球黑球分别有几个？概率有多少？贝叶斯则反向提出了问题：假设已知随机取出的 4 个球中有 3 个黑球，1 个红球，请问口袋 A、B 中红球、黑球分别是几个？概率有多高？这就是著名的贝叶斯“逆概”问题。不仅如此，18 世纪贝叶斯算法诞生两三百年以来科学家们发现贝叶斯定理颠覆了传统数学的很多观点，比如在对人脑思维模式的模拟过程中，人们一直认为计算机的体系与人脑的体系是完全两个不同的逻辑与算法系统，计算机只是凭借其“穷尽”的蛮力来产生“智能”，而人脑对单个事实或碎片的组织能力、推演能力是机器无法比拟的，这一切古老的认识在量子力学产生后发生了改变。在量子力学著名的电子“双缝干涉实验”中，电子通过左边还是右边完全是一个随机的过程，有时可能既是左边又是右边，完全取决于观测者当时的状况，就像薛定谔的猫，又是死又是活的叠加态。随机世界的产生为贝叶斯理论大行其道，大开方便之门，当代的人工智能探索中，贝叶斯概率分布已经完全模拟人脑的归纳总结能力，“今天晴转多云，降水概率 65%”，“这个病人手术后 5 年生存概率 51%”等。概率论是对随机事件最好的钥匙。

本案例中，应用本体论，贝叶斯网络算法对高血压患者心血管风险进行分类是对贝叶斯概率分布最有效的医学运用之一。本案例的主要思路是在确定病人患高血压的概率后，进一步推理出该病人由高血压引起的症状或者病症的概率，由随机理论推演风险事件的概率分布。其中变量的选择，父项、子项的选择与贝叶斯网络的构建都很有特色。最具亮点的特色是高血压知识库的构建，用计算机能懂的语言判断高血压患者的心血管风险概率分布是数据挖掘最重要的步骤，换言之，知识库是一种面向主题的知识集合，就是用机器语言预制逻辑推断所需要的专业知识，用标准化后的知识库来对逻辑判断提供对照与判断的依据。从手工计算的经典统计学到机器学习的典型数据挖掘，临床医学数据处理迎来了划时代的革新。既然机器也可以判断高血压的指征与诊断金标准，既然诊断金标准也可以标准化、定量化、模块化，机器智能模拟人脑的功能完全是可行的。

第 4 章

临床医学的模式识别

- 模式识别是什么
- 基线静息心率的故事
- 决策树算法
- 最大期望（EM）算法
- 算法的规律与临床医学的本质

4.1 模式识别是什么

4.1.1 定义

人们在观察事物或现象的时候，常常要寻找它与其他事物或现象的不同之处，并根据一定的目的把各个相似的但又不完全相同的事物或现象组成一类。模式识别又常称作模式分类，从处理问题的性质和解决问题的方法等角度，模式识别分为有监督的分类（Supervised Classification）和无监督的分类（Unsupervised Classification）两种。二者的主要差别在于，各实验样本所属的类别是否预先已知。一般说来，有监督的分类往往需要提供大量已知类别的样本，但在实际问题中，这是存在一定困难的，因此研究无监督的分类就变得十分有必要了。

模式还可分成抽象的和具体的两种形式。前者如意识、思想、议论等，属于概念识别研究的范畴，是人工智能的另一研究分支。我们所指的模式识别主要是对语音波形、地震波、心电图、脑电图、图片、照片、文字、符号、生物传感器等对象的具体模式进行辨识和分类。

临床医学的模式识别形式多样，一张 CT 片的判读，一个病理分型的确认，一种术式的疗效，一个靶向治疗的方案，临床医学的模式识别与疾病类型、生物特性、诊疗方案息息相关。临床医学的模式识别的类型主要有解释性建模、描述性建模、预测性建模、序列模式建模、知识性建模、依赖关系建模、异常与趋势建模等。

4.1.2 临床医学模式识别的故事

案例一：如何用模式识别的方法找到可预测结直肠癌预后的转录因子

近 10 多年来，由于辅助化疗的新药开发和新方案设计，并开展了一系列大规模临床研究，结直肠癌辅助化疗取得了长足进步，并逐渐改变了人们的一些传统观念。但这些成功的辅助治疗手段大部分都是针对晚期（Ⅲ期或Ⅳ期）结直肠癌，并不能改善早期（Ⅰ期或Ⅱ期）患者的无疾病生存期。

出现上面这种情况的原因在于目前还没有一种简单可靠的标准方法诊断高危复发的早期结直肠癌。为了这解决这个问题，研究人员尝试依据结直肠癌肿瘤组织的基因表达谱来对患者进行分层，并且目前已经开发出了可以识别高危结直肠癌的多基因表达标记芯片。

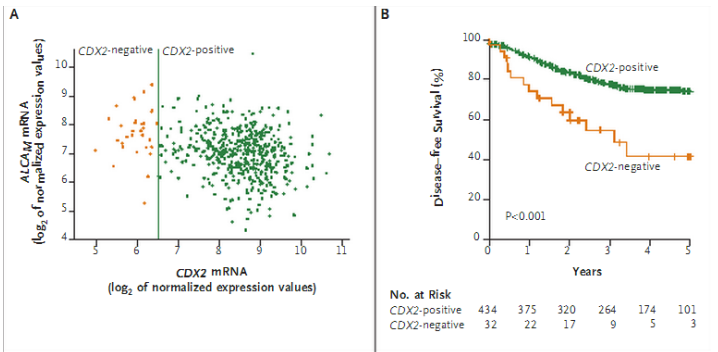
尽管基因表达标记芯片很有希望，但很难运用于临床，且不能预测辅助化疗的效果。在所有基因表达标记物中，来源于干细胞和祖细胞的标记物最有希望。因此，研究人员希望通过某种方法可以寻找这么一种标记物——能够识别未分化肿瘤。

为了寻找这种生物标记物，来自哥伦比亚大学的研究人员通过布尔逻辑生物信息学方法做了各种尝试，并最终发现了转录因子 CDX2，而且发现该标记物可以预测早期结肠癌预后，该研究最近发表在《新英格兰杂志》上。

该研究共纳入了 2115 例肿瘤标本，其中 87 例没有 CDX2 表达。466 例发现数据集中，32 例（6.9%）CDX2 阴性结肠癌患者 5 年无疾病生存期较 434 例（93.1%）CDX2 阳性结肠癌患者短。314 例验证

数据集中，38 例（12.1%）CDX2 蛋白阴性结肠癌患者 5 年无疾病生存期较 276 例（87.9%） CDX2 蛋白阳性结直肠癌患者短。该研究结果与患者的年龄、性别以及肿瘤的分级和分期无明显关系。

对于Ⅱ期结直肠癌，无论在发现数据集还是验证数据集中，CDX2 阳性和 CDX2 阴性患者的 5 年无疾病生存期都有明显差异。对所有患者的数据进行汇总分析发现，CDX2 阴性的早期（Ⅱ期）结直肠癌患者接受辅助化疗后的 5 年无疾病生存期明显较高。



C

Subgroup	Univariate Analysis		Multivariate Analysis	
	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value
All patients (N 466)				
CDX2-negative	2.71 (1.57 4.67)	<0.001	2.73 (1.58 4.72)	<0.001
Tumor stage, per increase in stage	3.47 (2.62 4.59)	<0.001	3.49 (2.61 4.67)	<0.001
Age, modeled as a continuous variable	0.99 (0.97 1.00)	0.058	0.99 (0.98 1.01)	0.37
Male vs. female sex	1.07 (0.89 1.28)	0.49	1.06 (0.88 1.28)	0.54
Patients with samples annotated with grading information (N 216)				
CDX2-negative	2.83 (1.42 5.64)	0.003	3.44 (1.60 7.38)	0.002
Tumor stage, per increase in stage	3.13 (2.14 4.60)	<0.001	3.28 (2.15 4.99)	<0.001
Tumor grade, per increase in grade	1.63 (0.94 2.82)	0.08	0.99 (0.56 1.74)	0.96
Age, modeled as a continuous variable	0.99 (0.97 1.01)	0.20	0.99 (0.97 1.01)	0.46
Male vs. female sex	1.15 (0.88 1.51)	0.32	1.20 (0.89 1.61)	0.24

图 4-1　CDX2 数据处理图表

本文为《良医汇：肿瘤资讯》原创，编译自：Biomarker in Stage II and Stage III Colon Cancer。

这也是一个解释性建模的案例，其本质还是模糊计算与建模。在辅助化疗技术对晚期癌症有效果的进步中鲜有对早期癌症有效的报道。本例中采用生物标记法成功得出 CDX2 阴性的早期（Ⅱ期）结直肠癌患者接受辅助化疗后的 5 年无疾病生存期明显较高。

案例二：肿瘤复发是死灰复燃，还是老根新芽？

来自美国、加拿大等数十个研究机构组成的一个国际研究小组发现的证据表明，在髓母细胞瘤（Medulloblastoma）患者体内复发的癌症与原发肿瘤并不一定是同一来源。这项研究结果发表在近日的《自然》（Nature）杂志上，该研究描述了他们利用小鼠模型开展的遗传工作，以及对人类肿瘤进行的全基因组测序，他们的研究发现且研究人员认为要更好地治疗患者需要去做更多的事情。

研究人员了解在手术切除首次发病的肿瘤后，形成的肿瘤在遗传上与原发肿瘤是否相同。由于在手术后都是基于原发肿瘤的遗传谱来量身定制设计治疗方法对抗新肿瘤生长，这一研究具有重要

的意义。为此，研究人员利用了睡美人转座子系统在测试小鼠中诱导髓母细胞瘤形成。他们随后通过模拟给人类患者处方的典型治疗进程对此进行了追踪——大约 60%的人类患者会出现肿瘤复发。研究人员从原发及复发肿瘤处采集组织样本后，进行了测序。

最终测序结果令人吃惊，在 11 个原发肿瘤中鉴别出了 23 个相同的插入位点，在复发肿瘤中鉴别出了 40 个相同的插入位点，然而，原发肿瘤与复发肿瘤之间的插入位点“极其得不同”。该研究小组随后对小规模的人类原发和复发肿瘤进行了全基因组测序。他们报告称在原发和复发肿瘤之间发现了“惊人”的遗传差异。这提示原发肿瘤与复发的肿瘤可能是完全不同的克隆来源。

研究人员由此认为，复发的髓母细胞瘤并非是原发肿瘤重新生长的结果，而是有可能由未被治疗靶向的一种不相关的子瘤生长而来。他们建议未来临床试验包括进一步的活组织检查应注意鉴别采用不同疗法靶向的子瘤。

资料来源：MedSci 转引自生物治疗科学网。

这是一个基因测序的模式识别案例，也是一个序列模式的模式识别。最重要的 KDD 是发现肿瘤的复发现象与原发肿瘤的重新生长无关，这类型肿瘤的复发可能是由未被治疗靶向的一种不相关的子瘤生长而来，这个模式识别的案例颠覆了经典的肿瘤复发学说。

案例三：CA125 对先兆流产结局的预示作用

背景：先兆流产影响 1/5 的女性，它与情绪困扰有显著相关性。先兆流产的预后仍然不能确定，由此成为了医疗护理专业人员的一项挑战。过去几年已经研究出多种生物标志物以预测先兆流产的结局；然而，结果却十分矛盾。因此，研究人员构建了一项系统性回顾性 meta 分析，以研究生物标志物对预测先兆流产女性妊娠结局的诊断准确性。

方法：这是一项系统回顾性 meta 分析的前瞻性研究，探究生物标志物预测孕龄在 5～23 周范围的女性出现先兆流产的预后。从 2015 年 6 月开始检索电子数据库，使用 QUADAS-2（诊断准确性的质量评估研究-2：修改工具）作质量评估，以计算诊断准确度。使用 Cochrane 系统评价软件做统计分析。

结果：共计 19 项研究被囊括到此次定性数据合成分析中，其中 15 项符合此次 meta 分析的纳入标准（包括 1263 例女性）。此次着重回顾以下生物标志物在预测先兆流产结局的作用：血清孕酮、hCG、与妊娠相关的血浆蛋白 A、雌二醇和 CA125。有趣的是，在胎儿存活率被发布后，发现血清 CA125 是最有潜力的标志物（七项研究的 648 例女性），而血清孕酮和 hCG 显得作用较弱。数据总结者在处理 CA125 的特征性时发现：敏感性为 90%（95%置信区间 83%～94%），特异性 88%（95%CI 79%～93%），正似然比 7.86（95% CI 4.23～14.60），负似然比 0.10（95% CI 0.06～0.20）。逆负似然比 9.31（95% CI 5～17.1），说明一个阴性的检测结果可以鉴别那些有继续妊娠可能性的群体。血清雌二醇是仅次于 CA125 的标志物，敏感性 45%（95% CI 6%～90%），特异性 87%（95% CI 81%～92%），正似然比 3.72（95% CI 1.01～13.71），负似然比 0.62（95% CI 0.20～1.84）。

结论：先兆流产女性的血清 CA125 水平在鉴定是否可以继续妊娠中具有高预测价值，而最常使用的血清 hCG 和孕酮在预测妊娠胎儿存活的结局时，显得并不理想。另外其他标志物，如抑制素 A

和标志物的结合需要进一步研究，希望可以在预测先兆流产女性妊娠结局方面有所发展。

原始出处：Pillai, R. N., et al. (2015). "Role of serum biomarkers in the prediction of outcome in women with threatened miscarriage: a systematic review and diagnostic accuracy meta-analysis." Hum Reprod Update.

资料来源：转引自 MedSci.

这也是一个预测性建模的故事。荟萃分析了 19 项研究的数据。先兆流产与情绪的高度相关已经为医学界所接受，在先兆流产预后很不确定的情况下采用荟萃分析的回顾性研究发现了 CA125 对预后的特异性，颠覆了传统医学认为的血清孕酮和 hCG 的作用，结论是先兆流产女性的血清 CA125 水平在鉴定是否可以继续妊娠中具有高预测价值，而最常使用的血清 hCG 和孕酮在预测妊娠胎儿存活的结局时，显得并不理想。这样的模式识别对临床、对患教都有极高的价值。传统的化验单只是注重 HCG 与孕酮指标，真相往往是在 HCG 与孕酮指标很好的笑逐颜开中危机正在来临，先兆流产的危险已经来到我们却全然不知。CA125 在模式识别中被发现的故事会改写医院的化验项目单。

4.2 基线静息心率的故事

静息心率又称为安静心率，是指在清醒、不活动的安静状态下，每分钟心跳的次数。早在 20 多年前，Framingham 研究就发现静息心率与心血管事件和死亡相关。然而多年过去，尽管后续有多项研究均显示心率加快与高血压、冠心病、急性心肌梗死及慢性心力衰竭等疾病的发病率和死亡率相关，但静息心率仍未被纳入主要心血管危险因素中，其中部分原因可能是受到其他危险因素的相互影响，另一原因可能是静息心率与心血管事件关系的机制仍未明确。已有研究发现，静息心率介导的动脉压与交感神经过度兴奋、动脉粥样硬化、斑块不稳定性有关，使其在心血管疾病进展和临床表现的潜在机制备受瞩目。众所周知，静息心率的加快不仅与心血管疾病的死亡率增加有关，同时也与非心血管疾病的死亡率相关。在既往许多流行病学研究和研究的冠心病患者中，静息心率与左心功能不全和（或）心力衰竭有关。这些与心率有关的现象都使得临床数据挖掘的方向指向这个值得研究的领域。

在许多假设机制中，心率升高可能直接影响心血管风险，多数与心肌需氧增加、能量缺乏、动脉粥样硬化进展或斑块破裂风险升高有关。

如果上述假设为真，心率数据隐含着什么真相？心率与哪些风险相关？心脏病患者的风险临界点能否用单一心率指标作风险提示？心率升高与结局之间的定量关系如何描述？能否用数据挖掘的手段给出答案？

如表 4-1 所示，所有 5438 例病例用基线静息心率 70 次/min 作为基线分割线，大于等于 70 为一组，小于 70 为另外一组。变量采用了四个大类（人口统计学特征变量 4 个，其中吸烟变量的加入是因为这个变量对心血管事件有特异的敏感性。既往病史变量 7 个，心脏参数 7 个，随机分组时的药物 8 类）。变量在空间中可以看成是一个个的维度，变量越多，维度越多，数据立方体（三维空间）的粒度越细。大数据条件下，粒度越细，维度越多有时候会造成“维度灾难”，这时候对计算机资源

占用过大，解决方案只能是用 N+1 维的升级空间维度的做法来缓解。就 P 值而言，性别、卒中史、血脂异常史、高血压病史四个指标没有显著的统计学意义，这是对原定假设的否定，就是说在这一轮的假设检验中有四个变量对于心血管风险没有意义。

表 4-1　患者基线特征

	心率<70 次/min（*n*=2745）	心率⩾70 次/min（*n*=2693）	*P*
人口统计学特征			
年龄（岁）	65.6（8.2）	64.4（8.6）	<0.0001
性别（男）	2298（84%）	2209（82%）	0.098
当前吸烟者	353（13%）	481（18%）	<0.0001
体重指数（kg/m^2）	28.3（4.1）	28.7（4.7）	0.0016
既往病史			
高血压病史	1911（70%）	1927（72%）	0.12
糖尿病史	846（31%）	1155（43%）	<0.000 1
血脂异常病史	2155（79%）	2123（79%）	0.77
既往心肌梗死	2468（90%）	2349（87%）	0.0019
经皮冠状动脉介入或冠状动脉搭桥术	1464（53%）	1360（51%）	0.037
卒中史	468（17%）	503（19%）	0.12
周围动脉疾病史	346（13%）	402（15%）	0.013
心脏参数			
心率（次/min）	64.1（2.8）	79.2（8.7）	…
收缩压（mmHg）	127.2（15.2）	128.5（15.7）	0.0017
舒张压（mmHg）	76.7（9.2）	78.3（9.2）	<0.000 1
左室射血分数（%）	32.7（5.3）	31.9（5.7）	<0.000 1
NYHA 心力衰竭分级 I 级	467（17%）	373（14%）	<0.000 1
NYHA 心力衰竭分级 II 级	1744（64%）	1615（60%）	
NYHA 心力衰竭分级 III 级	534（19%）	705（26%）	
随机分组时的治疗药物			
阿司匹林或抗血小板药物	2596（95%）	2507（93%）	0.023
血管紧张素转换酶抑制剂和（或）血管紧张素 II 受体抑制剂	2452（89%）	2421（90%）	0.049
β 受体阻断剂	2465（90%）	2273（84%）	<0.000 1
他汀类药物	2087（76%）	1945（72%）	0.0014
利尿剂（除外醛固酮拮抗剂）	1490（54%）	1704（63%）	<0.000 1
硝酸酯类药物	1133（41%）	1202（45%）	0.0123
醛固酮拮抗药物	666（24%）	800（30%）	<0.000 1

除非特别说明，数据以 *n*（%）或 $\bar{x}$（*s*）表示。NYHA=纽约心脏协会。1 mmHg≈0.133kPa

本研究的中心意义在于试验单一心率指标是否可以成为风险预警的阈值。

如图 4-2 所示，A、B、C、D 四个事件分别为心血管死亡、心力衰竭入院、心肌梗死入院、冠状动脉血运重建术。

结论一：在冠心病和左室收缩功能障碍的患者中，心率升高（≥70 次/min）意味着心血管疾病结局风险升高，同时对心力衰竭相关性结局与冠状动脉事件相关性结局有着不同的影响。

结论二：单一基线静息心率指标可以作为心血管风险事件的预警值。

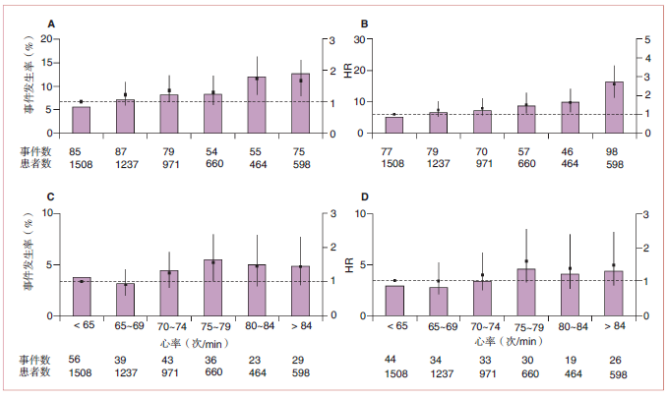


图 4-2　按照心率分组的事件粗发生率及 HR

按照心率分组的事件粗发生率（条；左侧标尺）以及 HR（95%CI，右侧标尺），与心率<65 次/min 有关的（A）心血管死亡、（B）心力衰竭入院；（C）心肌梗死入院；（D）冠状动脉血运重建术。水平线为 HR=1。

4.3 决策树算法

1. 案例

（1）背景

肝功能衰竭是肝癌肝部分切除术后危险和致命的并发症，为了减少手术风险，术前准确客观评估肝功能及预测术后剩余肝实质储备功能至关重要。吲哚氰绿（Indocyaniane green, ICG）是一种对人体无毒性仅在血管内分布的水溶性染料，注入血管后可高选择性地被肝细胞摄取，以游离形式由胆汁排出，且无肝肠循环，其排泄的快慢取决于肝细胞的功能，故可用 AEI 吲哚氰绿清除试验评价肝脏的储备功能。本研究回顾性分析 2009 年 2 月至 2010 年 7 月南京医科大学第一附属医院肝脏外科，采用东京大学肝胆胰外科制定结合 AEI 吲哚清除试验的决策树评估 82 例肝癌患者的肝功能而选择手术方式，探讨该决策树在评估肝脏储备功能中的临床应用价值。

（2）对象与方法

收集 2009 年 2 月至 2010 年 7 月在南京医科大学第一附属医院肝脏外科手术切除的 82 例肝细胞肝癌患者的临床资料。所有病例经组织病理证实均为肝细胞肝癌，并排除梗阻性黄疸如胆石症、胆道癌栓等胆道疾病。82 例肝细胞肝癌患者中男性 72 例，女性 10 例，年龄为 24~68 岁，中位年龄 46 岁。肝部分切除术采用钳夹法和超声外科手术吸引器（CUSA）。

ICGR-15 应用日本光电工业株式会社研发的 DDG-3300K 分析仪及相应的系统分析软件测定。注射用 AEI 吲哚氰绿由辽宁丹东医创药业有限责任公司生产（25 mg/支）。同时测量患者的血红蛋白、身高及体重。用灭菌用水配制 ICG 溶液（5 mg/dl），以 0.5 mg/kg 计算 ICG 药量静脉注射测量吲哚氰绿 15 min 储留率（ICGR-15）。 ICGR-15 的测定由南京医科大学肝癌研究中心完成。东京大学肝胆胰外科建立 1 个基于 3 个变量的决策树评估肝部分切除术的安全性：腹水、血清总胆红素水平和 ICGR-15。肝功能衰竭采用“50-50 标准”，即在术后第 5 天凝血酶原指数<50%（即 INR>1.7）及血清总胆红素>50μmol/L 同时存在。

如图 4-3 所示，应用 SPSS 10.0 统计软件进行分析，正态分布数据采用方差分析，计数资料采 X 方检验，$P < 0.05$ 认为有统计学意义。

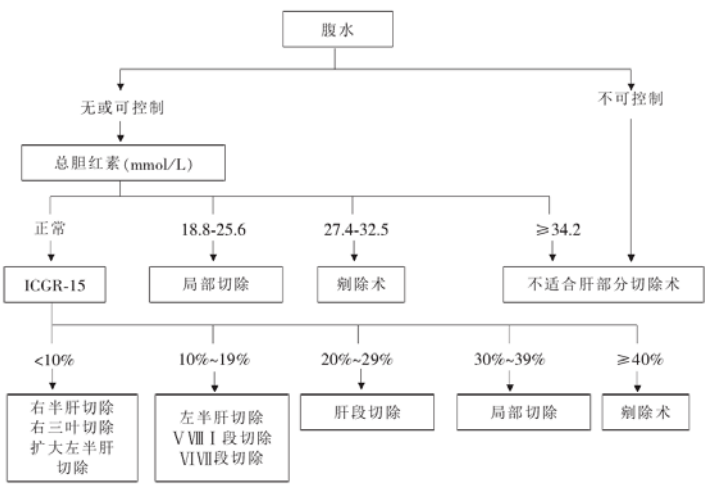


图 4-3　东京大学肝胆胰外科制定的结合吲哚氰绿清除试验的决策树

资料来源：吴晓峰等. 结合吲哚氰绿清除试验的决策树评估肝癌切除术患者肝储各功能临床应用. 南京医科大学学报. 2010 年 12 月.

本研究发现，82 例肝癌患者 ICGR-15 值随着 Child-Push 分级的递增逐渐升高，Child-Push A 级、B 级及 C 级患者 ICGR-15 分别为（7.8±3.2）%，（15.1±5.2）%和（22.8±6.1）%，3 组间比较差异有显著性（P<0.05），本组病例中通过术后病理证实 76 例合并结节性肝硬化（92.7%），合并结节性肝硬化患者 ICGR-15 平均为（17.5±3.2）%，明显高于无肝硬化患者的（6.5±1.8）%，P< 0.01。本

研究 82 例肝细胞肝癌患者中，总胆红素值 18.8～25.6 mmol/L 的 28 例患者行局部切除术，总胆红素值 27.4～32.5 mmol/L 的 12 例患者行肿瘤剜除术，其余总胆红素正常的 42 例患者 ICGR-15 平均值(18.5±3.7)%，17 例 ICGR-15 小于 10%[平均(6.3±2.1)%]患者行右半肝切除术和扩大左半肝切除手术；11 例 ICGR-15 值 10%～19%[平均(15.3±4.2%)]患者行 V VIQ 段切除和 VI VII 段切除术；8 例 ICGR-15 值 20%-29%[平均(24.3±3.5) %]患者行肝段切除术；4 例 ICGR-15 值 30～39%[平均(35.2±4.8)%]患者行肝局部切除术；2 例 ICGR-15 值>40%患者行剜除术。术后予保肝、营养支持及防治感染等治疗。术后并发症胆漏 2 例，可控制的腹水 9 例，少量胸水 8 例，切口愈合不良 6 例，经过积极的治疗后均治愈。本研究严格遵照东京大学肝胆胰外科制定的结合 AEI 吲哚氰绿清除试验的决策树行肝部分切除，术后无重大并发症及再次手术，所有患者包括 2 例 ICGR-15>40%的行肿瘤剜除术患者无 1 例发生肝功能衰竭及手术后死亡，术后均安全出院。

（3）重点讨论

肝储备功能指肝脏耐受手术及损伤的额外潜能，我国 85%以上肝癌患者都存在肝硬化基础，肝储备功能不同程度受损。肝储备功能较差的患者易在部分肝脏切除、麻醉、出血及手术打击后出现肝功能衰竭甚至死亡。目前对于肝部分切除术术后肝功能衰竭尚无统一定义，本研究采用“50-50 标准”，该标准是 Balzan 等研究发现，肝部分切除后患者在术后第 5 天凝血酶原指数<50%(即 INR>1.7) 及血清胆红素>50 μ N.,mol/L 同时存在，预测肝功能衰竭的发生率几乎为 100%，死亡率为 50%，是患者术后发生肝功能衰竭和死亡的准确预测因子。

目前广为使用的 Child-Push 分级是一种简便实用的半定量肝功能储备评估方法，但 Child-Push 分级法过于粗略，存在区分力低、敏感性差、主观性强及治疗干扰的缺点，因此只能初步评估肝脏功能状态，不能满足现代精准肝脏外科手术的实际需要。在亚洲如日本、中国香港等地区 ICG 是评估肝脏储备功能常用的方法，在中国大陆及欧美也开始应用 ICG 评估肝脏储备功能，ICGR-15 被认为能够客观精细地评价肝储备状况，是测定肝脏储备功能的理想方法。相对于 Child-Push 分级，ICGR-15 提供更多肝功能的信息，Child-Push 分级虽过于粗略，但此分级为临床应用的经典方法，结合 ICGR-15 指标可更完善评估肝功能储备状况。

香港范上达最初认为 ICGR-15<14%可耐受大部分肝脏切除手术，术后无严重并发症，目前普遍认为在年轻且有足够切除后残余肝脏体积的患者，ICGR-15 值可以推至 17%。日本东京大学肝胆胰外科幕内雅敏等认为在部分选择的患者行安全肝部分切除手术的 ICGR-15 上限为 40%。幕内雅敏等建立了一套用于术前安全切除范围的临床综合评估体系，该体系基于 3 个变量的决策树评估肝部分切除术的安全性：腹水、血清总胆红素水平及 IC-GR-15 值。总胆红素值升高(大于 2 mg/dl 即大于 34.2 mmol/L)或无法控制的腹水是肝部分切除术的禁忌症；总胆红素值 18.8～25.6 mmol/L 的患者行局部切除术；总胆红素值 27.4～32.5 mmol/L 的患者行肿瘤剜除术。总胆红素小于 17.1 mmol/L 的患者可根据 ICGR-15 值来决定切除范围，即 ICGR-15 小于 10%的可行肝脏体积的 2/3 的切除如右半肝切除、右三叶切除及扩大左半肝切除；10%～19%的可行肝脏体积的 1/3 切除术如左半肝等手术；20%-29%的可选择肝脏体积的 1/6 切除术如肝段切除术；若超过 30%就只可行局部切除或剜除手术。应用该决策树行术前安全切除范围的临床综合评估，东京大学肝胆胰外科在 107 例连续肝癌患者及 10 年中

685 例肝癌肝部分切除病例死亡率为零。本研究严格遵照东京大学肝胆胰外科制定的结合 AEI 吲哚氰绿清除试验的决策树行肝部分切除术，术后无肝功能衰竭及死亡，与东京大学肝胆胰外科报道结果相仿，证明结合 AEI 吲哚氰绿清除试验的决策树是安全有效的。

资料来源：吴晓峰等. 结合吲哚氰绿清除试验的决策树评估肝癌切除术患者肝储各功能临床应用. 南京医科大学学报. 2010 年 12 月.

2. 决策树算法的基本原理

（1）树以代表训练样本的单个结点开始。

（2）如果样本都在同一个类，则该结点成为树叶，并用该类标记。

（3）否则，算法选择最有分类能力的属性作为决策树的当前结点。

（4）根据当前决策结点属性取值的不同，将训练样本数据集 tII 分为若干子集，每个取值形成一个分枝，有几个取值形成几个分枝。均针对上一步得到的一个子集，重复进行先前步骤，递归形成每个划分样本上的决策树。一旦一个属性出现在一个结点上，就不必在该结点的任何后代考虑它。

（5）递归划分步骤仅当下列条件之一成立时停止：

① 给定结点的所有样本属于同一类。

② 没有剩余属性可以用来进一步划分样本。在这种情况下，使用多数表决，将给定的结点转换成树叶，并以样本中元组个数最多的类别作为类别标记，同时也可以存放该结点样本的类别分布。

③ 如果某一分枝 tc，没有满足该分支中已有分类的样本，则以样本的多数类创建一个树叶。

4.4 最大期望（EM）算法

案例：易感基因 GRlk4 与精神分裂症的关联分析及单倍型推断 EM 算法的改进

（1）单倍型的基本概念

单倍型是多个紧密连锁的 SNP 位点上位于同一条染色体上的等位基因的集合，它含有多个位点间的连锁不平衡信息。大量研究表明把相关区域的 SNPs 作为一个整体来考虑，会得到更多的信息。因此，就基因定位等问题来说，基于单倍型的分析比基于单个 SNP 位点的分析有更大的功效。由于分子方法测定单倍型效率低、费用高，而目前的分型和测序技术并不能提供连锁相的信息，通过统计学的方法推断出个体的单倍型是目前最为有效的方法。EM 算法是最为常用的推断算法之一。为了克服标准 EM 算法的缺陷，研究者发展了多个基于 EM 的改进算法，如 PLEM、OSLEM、SEM 等。然而，这些程序包中的大多数是针对 SNP 数据的，也就是每个位点只能有两种等位基因的数据。

（2）EM 算法的改进

在复杂疾病的连锁和关联分析中，越来越多的研究需要同时分析多等位基因位点和 SNP 位点。因此，针对多等位基因位点的 LD 分析和单倍型推断是一个非常普遍的问题。由于上述情况，不少研究者最近又开发了新的程序，如 MIDAS 和 ISHAPE 等。在标准 EM 算法推断单倍型的过程中，最大的问题在于杂合位点较多的个体，会有很多种单倍型组合和基因型相匹配。而对于整个人群而言，

就需要估算数目巨大的单倍型的频率，加上由于计算机存储能力的限制，标准 EM 算法就不能处理太多位点。一个解决方法就是限制需估算频率的单倍型数目。那么，哪些单倍型的频率需要估算呢?构建一个合理的需估算频率的单倍型集合成为关键。一个好的候选单倍型集合有助于又快又准地估算出单倍型频率。在构建单倍型候选集合的过程中，一个比较合理的方式是逐步构造这个集合，而并非一步完成，考虑所有可能情况。本论文提出了一个新的算法 PL-CSEM（Pardon Ligation Combination Subdivision EM），该算法基于标准 EM 算法，结合了 Combination Subdivision（CS）策略，来提高算法处理多等位基因位点数据的能力，特别是针对各位点等位基因数较大的数据。此外，由于 Partition Ligation（PL）策略在处理多位点数据的有效性[45,52]，我们也把该策略应用到我们的算法中，形成了 PL-CSEM，使其在处理多位点及多等位基因位点的数据时有一定的优越性。

EM 算法是求参数极大似然估计的一种方法，它可以从非完整数据集中对参数进行 Maximum Likelihood Estimate（MLE）。广泛应用于处理缺损数据、截尾数据等不完全数据（Incomplete Data）[47,238]。既然对于含有多个杂合位点（大于或等于 2 个）的个体的单倍型的确定可以作为不完全数据来考虑，Excoffier 和 Slatkin 于 1995 年首先提出在假设 HWE 的前提下利用 EM 算法推断单倍型。

（3）CS 策略

CS 策略的具体步骤如下：对于多等位基因位点，首先把某些等位基因当作同一个中间等位基因来处理（Combination 步骤），这样，可以减少该位点的等位基因数目，从而可减少需考虑的单倍型数目。然后再用 EM 算法来进行单倍型分析。当然，这一步得到的这些单倍型是由那些中间等位基因构成的，这里称这些单倍型为中间单倍型。因而，需要 Subdivision 步骤，在这一步骤中，自上而下逐步分解那些 Combination 步骤生成的中间等位基因，直到这些等位基因和原始的一致为止。例如，我们需要对含有两个多等位基因位点（A 和 B）的数据进行单倍型分析（图 4-4），不失一般性地假设他们的等位基因数分别 k 和 l 位，那么可能出现的单倍型共有 kl 种。标准 EM 算法会把所有可能出现的单倍型一次性都纳入考虑，这里，它需要一次性考虑 kl 种单倍型。但是往往会因为需考虑的单倍型数目过多，如果这里采用的计算机存储能力有限，而使算法无法运行，或者耗费时间过长。而 CS 策略，对于每个位点，先把某些等位基因结合起来，等位基因位点当作双等位基因位点，然后再用 EM 算法来分析单倍型，需要考虑的中间单倍型仅有 4 种。然后，只有频率大于某一阈值的中间单倍型才会被保留进入下一步骤。在每一个 Subdivision 步骤中，那些中间等位基因都会被细分成两部分，从而，那些被留下的由它们构成的中间单倍型也会被分解，形成一个候选的单倍型集合。而后，再利用这个集合来进行 EM 推断。Subdivision 步骤会一直重复，直到所有的中间等位基因不再出现，也就是到推断出的所有的单倍型都是原始的等位基因构成，循环结束，所得到的结果就是我们想要的最终单倍型。

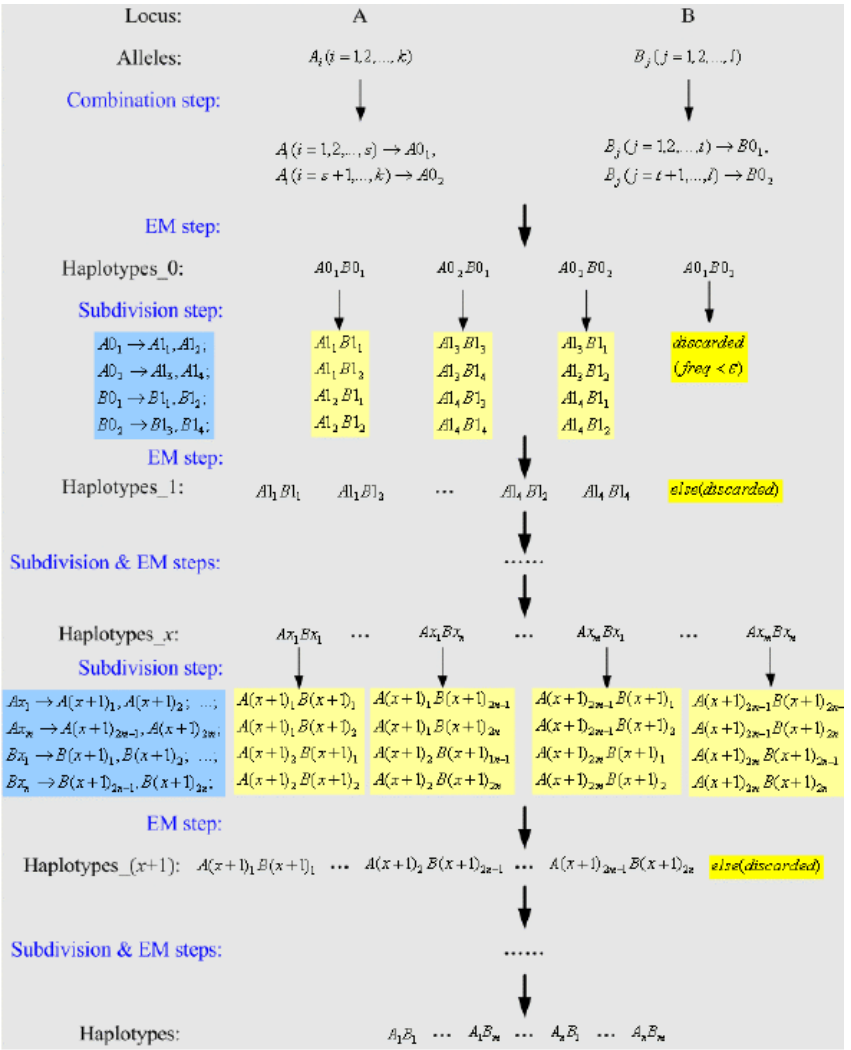


图 4-4 CS-EM 算法流程图

资料来源：李志强. 易感基因 GRIk4 与精神分裂症的关联分析及单倍型推断 EM 算法的改进. 上海交大. 2008 年硕士论文.

（4）PL 策略

考虑到一个多位点的单倍型在本质上是由多个小段结构组成的，PL 方法采用自下而上的策略，在处理多位点的过程中，首先把这些位点分成小块，每个小块由连续位点组成(典型长度是 5 到 8 个位点)，而后再把这些小块拼接起来形成整块的单倍型。2002 年，Qin 等人结合了 PL 策略和标准 EM 算法形成了 PLEM 算法，大致过程和上面提到的一样，不同是对于小块单倍型推断和拼接过程中所

采用的算法都是 EM 算法。不过，该程序只针对 SNP，不能处理含有多等位基因位点的数据。

如图 4-5 所示，由于 PL 策略能够有效处理多位点的数据，我们也应用了该策略。与 PLEM 不同的是，对于每个小块，如果其中包含多等位基因位点，我们用的是 CS-EM 并非 EM 来推断单倍型。此外，拼接过程与 PLEM 的也有不同(图 4-5)。PLEM 是把相邻的两个块逐步拼接，作为一个多等位基因位点来考虑，而这里采用了另一种方法。

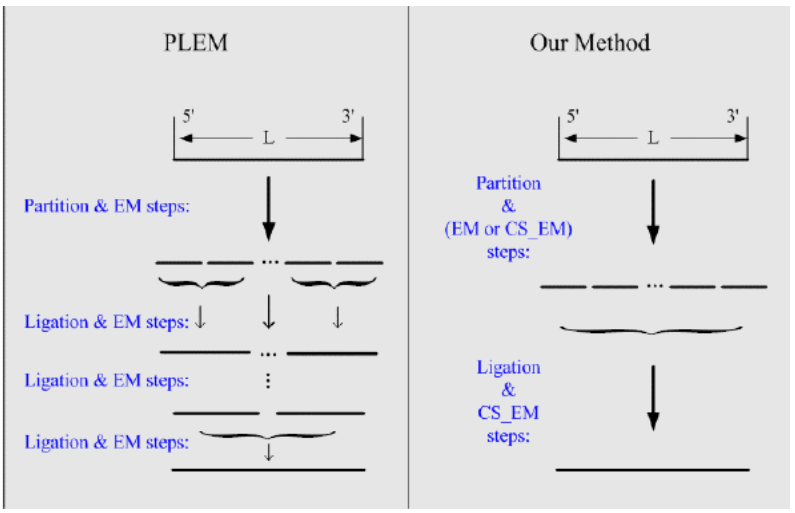


图 4-5　PLEM 和 PL-CSEM 的不同点

资料来源：李志强. 易感基因 GRIk4 与精神分裂症的关联分析及单倍型推断 EM 算法的改进. 上海交大. 2008 年硕士论文.

（5）性能分析

在所有参数的比较中，绝大多数情况下，PL-CSEM 的表现都好于 PLEM。由于 HLA 数据的真实单倍型数据无法获得，所以有关它的这些参数值无法计算。PL-CSEM 和 PLEM 程序对于真实数据的 I_H 和 I_F 情况如表 4-2 所示，对于模拟数据 I_F 和 I_H 情况如表 4-3 所示。

表 4-2　PL-CSEM 和 PLEM 程序的 I_H、I_F 数据

GH1		Chr12-HapMap_CEU	
15Loci		20SNPs	40SNPs
I_H			
PLEM	-	0.867	0.764
PL-CSEM	0.784	0.877	0.770
I_F			
PLEM	-	0.948	0.867
PL-CSEM	0.927	0.954	0.887

由于 PLEM 不能处理含有多等位基因位点的数据，所以上面部分值无法计算。

资料来源：李志强.易感基因 GRIk4 与精神分裂症的关联分析及单倍型推断 EM 算法的改进. 上海交大. 2008 年硕士论文.

4-3　程序的 I_H 和 I_F 均值对比

Sample Num	60		500				1000			
Loci Num	20	40	20	40	80	160	20	40	80	160
I_H										
PLEM	0.752	0.665	0.911	0.877	0.827	0.761	0.964	0.941	0.913	0.880
PL-CSEM	0.806	0.710	0.931	0.915	0.905	0.895	0.976	0.972	0.977	0.979
I_F										
PLEM	0.864	0.747	0.983	0.966	0.951	0.932	0.991	0.984	0.975	0.965
PL-CSEM	0.899	0.812	0.986	0.980	0.972	0.955	0.993	0.990	0.990	0.987

PL-CSEM 和 PLEM 程序运算这些测试数据的错误率情况如表 4-4、表 4-5 所示，图中 INDIVIDUALS 和 LOCI 分别表示基于个体和基于位点的单倍型推断错误率。

表 4-4　程序的单倍型推断错误率(%)均值对比情况(对于真实数据)

GH1	Chr12-HapMap_CEU	
15Loci	20SNPs	40SNPs
INDIVIDUALS		
PLEM	-	4.55	13.23
PL-CSEM	9.09	4.00	11.32
LOCI		
PLEM	-	0.39	0.81
PL-CSEM	0.82	0.36	0.75

INDIVIDUALS 和 LOCI 分别表示基于个体和基于位点的单倍型推断错误率，由于 PLEM 不能处理含有多等位基因位点的数据，所以上面部分值无法计算。

表 4-5　程序的单倍型推断错误率(%)均值对比情况(对于模拟数据)

Sample Num	60		500				1000			
Loci Num	20	40	20	40	80	160	20	40	80	160
INDIVIDVALS										
PLEM	11.28	24.25	1.08	2.61	3.70	4.13	0.60	1.29	1.97	2.05
PL-CSEM	8.05	18.15	0.08	1.57	2.73	5.97	0.49	0.81	1.47	2.75
LOCI										
PLEM	0.92	1.51	0.08	0.15	0.17	0.15	0.04	0.07	0.09	0.07
PL-CSEM	0.63	1.03	0.06	0.09	0.11	0.16	0.03	0.04	0.06	0.08

INDIVIDUALS 和 LOCI 分别表示基于个体和基于位点的单倍型推断错误率。

（6）结论

在现代生物遗传学中，生物体的单倍型信息起着非常关键的作用，可以提高连锁分析和关联分析的功效和准确度，是人类探究基因多样性以及定位复杂疾病基因的重要手段之一。随着分型和测序技术的发展，可以高效获得生物体的基因型数据，但是单倍型信息却不能直接得到。十几年来很多学者都致力于这方面的研究，发展了一系列通过统计学方法利用基因型数据推断个体的单倍型信息，如 Clark's 算法、EM 算法、Bayesian 方法和 PGS 方法等。EM 算法是其中应用较为广泛的算法之一，因为它有稳定及收敛速度较快等特性。研究者发展了多种 EM 算法的衍生算法，如 OSLEM、SEM 和 PLEM 等。然而大多数基于 EM 算法的程序包都是针对 SNP 的，很少有针对多等位基因位点的。即便能够处理多等位基因位点数据，也没有采用相应的策略，使其能够高效处理多等位基因位点的数据。

现实中，常常会遇到需要处理含有多等位基因位点的情况，因而，我们开发了 PL-CSEM 程序，该程序基于标准 EM 算法，同时结合了 Combination Subdivision 策略和 Partition Ligation 策略逐步构建候选单倍型集合，避免遇到需考虑的单倍型数目过多的问题，从而避免过大的存储消耗，减少运行时间。CS 策略针对的对象是多等位基因位点的等位基因个数，这一点不同于以往研究所采用的策略。考虑到多位点数据也很常见，以及 PL 策略在处理这类数据的有效性，和 PLEM 一样，我们的程序也采用了该策略，只是在某些时候用法不同，上面已有详述。

在我们测试的大量数据中，PL-CSEM 程序在绝大多数情况下表现好于 PLEM，具体表现在获得更高的 I_H 和 I_F，以及单倍型推断误差率更小(无论是基于个体的，还是基于位点的)，还有运行时间更短。当然，在个别情况下，PL-CSEM 程序的某些参数不如 PLEM 的，如样本量为 500 和 1000，位点数为 160 的情况下，PL-CSEM 程序的基于个体的单倍型推断错误率分别为 5.97%和 2.75%，略高于 PLEM 程序的 4.13%和 2.05%。此外，它对于样本量为 500，位点数为 160 的数据的运行时间也略多于 PLEM，但是对于样本量为 1000，位点数为 160 的数据，PL-CSEM 程序的运行时间却大大小于 PLEM 程序的，不及其 1/8。而且，实质上，PL-CSEM 程序的这个单倍型推断错误率也不是很大，基本上可以接受。

4.5 算法的规律与临床医学的本质

4.5.1 算法的本质是什么

利用关联规则的 Apriori 算法分析高考成绩对医学生基础医学课程成绩和临床医学课程成绩的影响，总结高考成绩对医学生专业课成绩影响的规律。这些规律对专业建设、培养方案制定、课程设置、教学效果检查、教学方法改进等有积极作用，并且能够为教学管理决策提供依据。这就是算法的本质：模式识别。

利用高考成绩的数据预测其对医学生专业课成绩的影响规律是一种典型的关联规则算法。这其

中变量的设置十分关键，比如这名同学的高考化学课成绩非常好，我们就可以用关联规则算法找出高考化学成绩好的同学在医学专业课上有什么结果？是否药理学成绩会有不俗的表现？这些关联性之间有什么规律？这就是算法的本质：找到关联性。

在很多的临床医学与生物医学的假设机制中，预设的假设条件与命题要经过检验来证伪或证实（*T* 检验，*P* 检验等），在数据中抽样，在参数设计后考虑方差检验，按最大的似然原则提取结果，这就是模式识别的本质：医学的知识发现。

无论是解释性建模、描述性建模、预测性建模、序列模式建模、依赖关系建模还是异常与趋势建模都可以具体表现为大数据的分类、回归分析、聚类、关联规则、神经网络方法，这些方法从不同的角度对数据进行挖掘，其本质是借助计算机从数据中找规律。

4.5.2　数据挖掘中医学的本质

数据挖掘中的医学呈现什么样的本质？我们通过案例来说明。

来自伦敦大学学院与牛津大学等机构的科学家们在近期《自然》（Nature）杂志上报告称，通过挽救控制微血管血流的周细胞（pericyte），有可能减轻中风引起的长期脑损伤。此前许多研究人员都认为，脑内血流量仅受微动脉直径变化控制。而这项最新研究表明，实际上主要是周细胞缩紧或松开毛细血管时，毛细血管扩大或缩小控制了大脑血液供应。

研究发现，周细胞不仅是大脑血流量的主要控制因子，中风之后它们还收缩并死在毛细血管的周围。这显著地损害了长期的血流量，对脑细胞造成持久的破坏。研究人员证实，某些化合物可以减少一半实验室中模拟中风导致的周细胞死亡，并希望能将这些化合物开发为药物来治疗中风患者。

论文资深作者、伦敦大学学院神经科学教授 David Attwell 表示：“如果中风患者能够及早到达医院，目前临床医生可以将堵塞大脑血流的血凝块除去。然而，周细胞引起的毛细血管收缩，通过长期限制血液供应，可在除去血凝块后进一步对神经细胞造成损伤。我们最新的研究表明，设计一些药物来阻止毛细血管收缩或许可为减轻中风引起的功能障碍提供一些新疗法。”

论文作者之一、牛津大学医学部主任 Alastair Buchan 教授表示：“这一发现提供了全新的中风治疗策略。重要的是，我们现在应该能够鉴别出一些靶向这些细胞的药物。如果我们可以阻止周细胞死亡，应该会有助于将大脑的血流量恢复至正常，阻止中风后我们看到的进行性发展的慢性损伤，避免造成患者的神经功能失常。”

此外，这项最新研究还提供了利用功能性磁共振成像来检测大脑血流量改变一些潜在机制的新认识。Attwell 教授表示：“功能性成像使得我们能够看到人类大脑中的神经细胞活动，但直到现在我们都并不是很清楚我们看到的是什么。我们证实，是周细胞启动了神经细胞活跃时看到的血流量增高，因此现在我们知道了，功能性的成像信号是由周细胞介导的毛细血管直径扩张所引起。确切地了解功能成像所显示的东西，将帮助我们更好地理解及注释我们所看到的一切。”

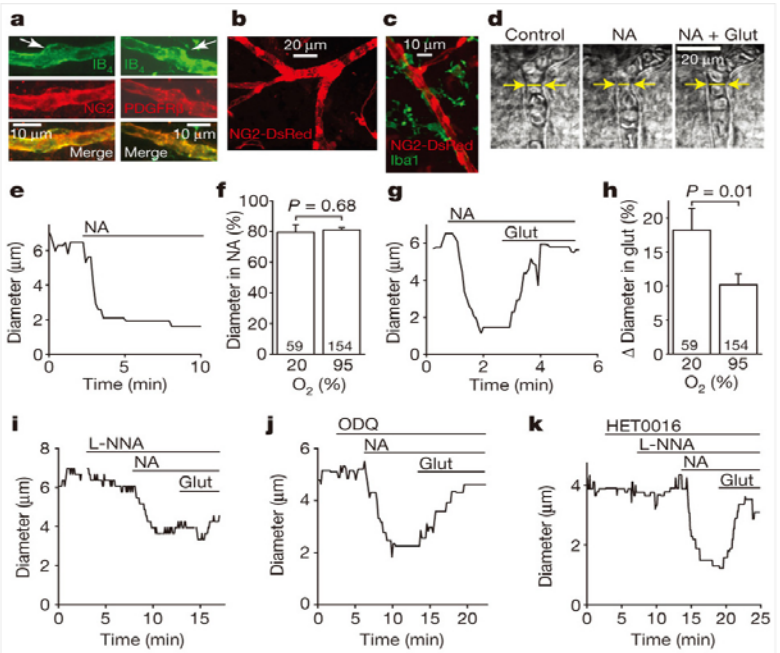


图 4-6　微血管血流的周细胞收缩特性的药物靶向控制机制

资料来源：“NATURE” Capillary pericytes regulate cerebral blood flow in health and disease Nature Volume: 508 Pages: 55–60 Date published:(03 April 2014)

如图 4-6（a、b、c、d、e、f、g、h、i、j、k）所示，动物白鼠实验中对微血管血流的周细胞（ericyte）收缩特性的药物靶向控制机制被一个个数据图表栩栩如生地表达出来。在分子级靶向标记药物导引下，我们看到了显微成像中微周细胞在舒张、收缩、受控状态下不同的直径大小与血流流量与走向的变化（图 a-d）。再比如图（f）有力地表现了在使用了去肾上腺素（NA）后微周细胞的直径大小，收缩程度与氧分子的含量无关，20%与 95%氧分子含量条件下的直径大小相差无几。图（j）则表达了微周细胞在不同的化合物与蛋白转运酶条件下不同的收缩直径与时间。在精准的化合物作用下，脑细胞中的微周细胞收缩可以得到控制从而改善了脑卒中后供血情况。这是一个改写脑中风临床指南与路径的重要发现，传统的理论认为微动脉血管是脑细胞主要供血路径的说法受到了挑战，是周细胞启动了神经细胞活跃时看到的血流量增高，对周细胞的挽救与唤醒可以极大地改善脑卒中患者的预后。本案例深刻地表达了医学的本质就是数据的解析，在分子层次对细胞作用机制的数据描述，对研究过程数据、图表、成像的全过程模拟，完美再现了数据诠释生物与医学现象的独特优势。在科学论文严格的范式下，数据技术与医学的结合相得益彰。

不难想象，数据可以重建全身的毛细血管，数据可以完全模拟蛋白转运酶的机制，数据当然可以诠释医学的原理。

我们再看一个案例。

背景：患者在多种常见手术后再住院的现象十分常见，但患者在原手术医院（导入医院）再住院并接受治疗后其结局是否有所改善仍不明确。本研究分析了美国 Medicare 受益者在接受多种常见手术后再住院与死亡风险之间的相关性。

方法：应用 2001 年 1 月 1 日至 2011 年 11 月 15 日间的美国 Medicare 受益者报销数据，本研究评估了行开腹腹主动脉瘤修复术、腹股沟下动脉旁路移植术、主动脉双侧股动脉分流术、冠状动脉旁路移植术、食管切除术、结肠切除术、胰切除术、胆囊切除术、腹疝修复术、开颅术、髋关节置换术或膝关节置换术后 30d 内需要再住院治疗的患者。采用包含逆概率加权的 logistic 回归模型和辅助变量分析在内的方法，对接受手术治疗后再住院患者的再住院（导入医院和非导入医院）与 90 d 死亡风险之间的相关性进行检测。

结果：9440503 例患者接受了 12 种主要手术之一，这些患者再住入或被转入导入医院的比例介于接受冠状动脉旁路移植术的比例 65.8%（186 336/283 131）与接受结肠切除术的比例 83.2%（142 142/170 789）之间。如果因手术并发症而再住院，在导入医院的再住院比例高于非导入医院[189 384/834 070（23%）和 36 792/276 976（13%），$P<0.000\ 1$]。采用逆概率加权控制选择偏倚后，在导入医院再住院患者的 90 d 死亡率较非导入医院降低 26%（OR 0.74，95%CI 0.66～0.83）。逆概率加权模型分析显示，该效应在本研究纳入的所有类型手术中均具有显著性（$P<0.000\ 1$），并且在胰切除术（OR 0.56，95% CI 0.45～0.69）与主动脉双侧股动脉分流术（OR 0.69，95%CI 0.61～0.77）中最具显著性。在辅助变量分析中，采用区域导入医院再住院率中的医院水平变量作为辅助变量，在导入医院再住院概率最高的患者较导入医院再住院概率低的患者的死亡风险降低了 8%（OR 0.92，95%CI 0.91～0.94）。

结论：在美国，接受多种常见手术后的患者在导入医院再住院治疗后，其存活率有所改善。这些结果也许对手术治疗的成本-效益驱动型区域集中化有重要意义。

资料来源：《柳叶刀中文版》 2015 年 第 9 卷 第 10 月 第 10 期

本案例再一次有力地证明了数据挖掘技术的重要性，导入了医学中常用的逆概率加权法。在许多长期随访，且以死亡、肿瘤进展作为观察结局的临床试验中，如果研究的中期评价发现新药在某些方面（如无病进展时间）显著的优于对照药，按照伦理学的规定，对照组的患者将可以在之后的治疗中选择改变治疗方案，接受新药治疗。这种“选择性的治疗转换”破坏了随机化原则，导致试验中对象的生存时间和删失时间不独立。对于这样的背景问题，利用 Cox 模型考察组间的死亡风险比时，可能得到有偏的参数估计。另一方面，在观察性研究中，若存在一个随时间变化的协变量，是研究的混杂因素，且会受到前次暴露的影响，此时常用的分析方法在估计暴露效应时也是存在偏倚的。以上两个问题虽然来自临床试验和纵向数据两种背景，然而都可以通过逆概率加权的分析方法加以解决。除了逆概率加权法外，本案例令人印象深刻的就是大数据量的病案数据分析，长达 11 年的 940 万个案例数据，横跨 12 个手术病种，有力地论证了再入院对手术预后的影响。这也是数据挖掘中医学的本质：在数据中找到真相与规律，在数据中找到手术的自信。

小知识：算法是什么？

“如何更快地叫到出租车”、“如何在社交网站找到称心的伴侣”、“如何判断患者是否罹患癌症”、“如何判断一笔交易是否属于欺诈”……这些问题看上去风马牛不相及，究其根本，其实都有算法在幕后“操纵”。那么，什么是算法？常见的算法类型有哪些？算法跟我们的日常生活有什么关系呢？什么是算法？算法就是通过从数据里提取规则或模式来把数据转换成信息。目前通常所说的算法一般分为两类，一类是传统算法，就是算法导论一类的书里面介绍的算法，而另一种应该称作模型，比如神经网络、SVM、K-Means 等。常见的算法类型有哪些？

（1）分类算法

分类算法按照某种标准给对象贴标签，再根据标签来区分归类，主要包括逻辑回归、贝叶斯判别、随机森林等。

（2）预测算法

预测算法和分类算法最大区别在于，分类算法的目标变量是分类离散型（例如，是否逾期、是否肿瘤细胞、是否垃圾邮件等），预测算法是连续的（股票趋势的预测），包括线性回归、神经网络、SVM 等。

（3）聚类分析

聚类分析，通过距离，将所有样本划分为几个稳定可区分的群体，常见的聚类算法包括 K-Means、系谱聚类、密度聚类等。

（4）关联分析

关联分析，通过数据发现项目 A 和项目 B 之间的关联性，找出内在的联系。常常是指购物篮分析，即消费者常常会同时购买哪些产品（例如游泳裤、防晒霜），从而有助于商家的捆绑销售。上文所提到的四种算法类型（分类、预测、聚类、关联），是比较常见的。还有其他一些比较有趣的算法类型和应用场景，例如协同过滤、异常值分析、神奇的图像处理算法等。

邮箱系统如何分辨一封 Email 是否属于垃圾邮件？一般来说，判断邮件是否属于垃圾邮件，先是把邮件正文拆解成单词组合，然后根据贝叶斯条件概率，计算一封已经出现了某单词的邮件，属于垃圾邮件的概率和正常邮件的概率。如果结果表明，属于垃圾邮件的概率大于正常邮件的概率。那么该邮件就会被划为垃圾邮件。

一只南美洲热带雨林中的蝴蝶，偶尔扇动了几下翅膀，可以在两周以后，引起美国德克萨斯州的一场龙卷风。你在互联网上的搜索是否会影响公司股价的波动？一个公司在互联网中搜索量的变化，会显著影响公司股价的波动和趋势，即所谓的投资者注意力理论。该理论认为，公司在搜索引擎中的搜索量，代表了该股票被投资者关注的程度。因此，当一支股票的搜索频数增加时，说明投资者对该股票的关注度提升，从而使得该股票更容易被个人投资者购买，进一步地导致股票价格上升，带来正向的股票收益。

采用支付宝进行支付时，或者刷信用卡支付时，系统会实时判断这笔刷卡行为是否属于盗刷。通过判断刷卡的时间、地点、商户名称、金额、频率等要素进行判断。这里面基本的原理就是寻找异常值。如果您的刷卡被判定为异常，这笔交易可能会被终止。

异常值的判断，应该是基于一个欺诈规则库的。包含两类规则，即事件类规则和模型类规则。第一，事件类规则，例如刷卡的时间、地点、金额、频次等是否异常。第二，模型类规则，则是通过算法判定交易是否属于欺诈。一般通过支付数据、卖家数据、结算数据，构建模型进行分类问题的判断。

电商中的猜你喜欢，应该是大家最为熟悉的。在京东商城或者在淘宝购物，总会有“猜你喜欢”、“购买此商品的顾客同时也购买了**商品”，这些都是推荐引擎运算的结果。

一般来说，电商的“猜你喜欢”（即推荐引擎）都是在协同过滤算法的基础上，搭建一套符合自身特点的规则库。即该算法会同时考虑其他顾客的选择和行为，在此基础上搭建产品相似性矩阵和用户相似性矩阵。基于此，找出最相似的顾客或最关联的产品，从而完成产品的推荐。

第 5 章

医学数据挖掘的常用工具

- SAS 挖掘软件运用案例
- Weka 软件介绍
- Matlab 案例
- R 语言案例
- 临床医生如何用好挖掘工具

5.1 SAS 挖掘软件运用案例

（1）SAS 软件

SAS（全称 Statistical Analysis System，简称 SAS）是全球最大的软件公司之一，是由美国 NORTH CAROLINA 州立大学 1966 年开发的统计分析软件。1976 年 SAS 软件研究所（SAS INSTITUTE INC）成立，开始进行 SAS 系统的维护、开发、销售和培训工作。期间经历了许多版本，并经过多年来的完善和发展，SAS 系统在国际上已被誉为统计分析的标准软件，在各个领域得到广泛应用。SAS 软件如图 5-1 所示。

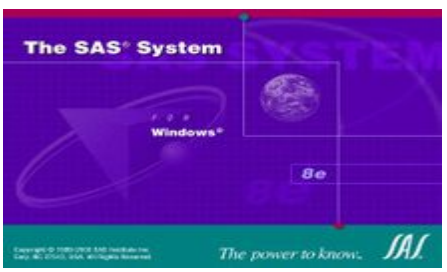


图 5-1　SAS 软件

（2）运用案例

在医学统计学中，为了客观、全面地分析问题，常要记录多个观测指标并考虑众多的影响因素，这样的数据虽然可以提供丰富的信息，但同时也使得数据的分析工作更趋复杂化。例如，现要对某地区 16 岁男孩的生长发育情况进行评价，通过调查收集了身高、坐高、体重、胸围、肩宽与骨盆宽 6 个指标（见表 5-1），显然，这 6 个指标间存在着相互联系和影响，要对其进行综合评价，可以采用主成分分析的方法。

表 5-1　某地区 16 岁男孩身体形态学指标

编　　号	身高（cm）	坐高（cm）	体重（kg）	胸围（cm）	肩宽（cm）	骨盆宽（cm）
1	165.13	88.76	51.60	78.31	35.39	25.96
2	164.83	89.87	53.32	83.04	37.24	26.52
3	164.58	88.15	51.38	80.49	36.05	26.27
4	164.56	87.02	51.18	81.63	36.14	26.84
5	164.46	86.84	54.52	81.21	35.67	26.74
6	164.13	87.18	50.14	81.88	36.88	26.13
7	164.64	87.00	49.85	78.99	35.45	26.00
8	163.64	86.89	50.85	82.09	35.54	26.01
9	163.45	86.99	52.28	79.04	35.78	25.51

续表

编　　号	身高（cm）	坐高（cm）	体重（kg）	胸围（cm）	肩宽（cm）	骨盆宽（cm）
10	163.00	88.11	50.48	80.37	36.43	26.87
11	162.40	86.58	47.46	79.38	35.50	24.65
12	162.18	86.95	52.18	82.50	36.85	26.43
13	162.04	86.71	48.93	78.00	35.59	25.07
14	161.86	87.35	49.03	79.69	35.68	25.22
15	160.78	84.92	47.99	77.76	35.49	25.64
16	160.77	85.35	48.70	81.63	35.58	26.92
17	160.69	86.18	48.94	81.89	35.47	26.81
18	160.44	86.05	47.05	78.88	35.13	25.71
19	160.27	86.41	51.51	79.84	35.66	26.02
20	159.05	85.08	48.95	80.87	35.46	25.85
21	168.25	85.14	49.31	79.32	36.02	25.59
22	158.18	84.05	46.93	79.83	34.74	25.26
23	157.98	84.49	47.60	79.04	35.11	25.66
24	157.61	84.57	43.39	78.45	35.20	26.15
25	157.23	83.90	46.90	78.59	34.73	25.57
26	157.05	84.55	46.54	78.61	35.20	25.90
27	153.13	81.63	43.40	77.42	33.82	24.62

资料来源：惠晓萍. SAS 软件在医学统计主成分分析中的应用. 中国现代医药杂志 2011 年 8 月第 13 卷第 8 期.

主成分分析是从多个数值变量之间的相互关系入手，利用降维的思想，将多个变量化为少数几个互不相关的综合变量的统计方法，其利用各因素的评价指标构造参数矩阵，求解参数相关矩阵的特征根及特征根对应的特征向量，根据给定的信息确定主特征根和主特征向量。因为由相关矩阵确定的主特征向量相互独立，且几乎包括了所有特征向量所代表的信息，所以称之为主成分向量。主成分分析的目的是用较少个数的综合指标来反映全部原始指标的主要信息。

编制 SAS 程序如下：

```
Title'16 岁男孩生长发育评价
data  principal;
input  number X1-X6@@;
cards;
……(表 1 数据)
proc
Var
princomp data=principal  prefix=Z
X1-X6;
```

```
Run;
proc
Var
print data=principal;
number Z1 Z2 X1-X6;
run;
```

运行结果：

单击 SAS 菜单栏中的 Locals→Submit，得到如 h 结果，如表 5-2 所示。

表 5-2　简单统计量

项　目	*X*1	*X*2	*X*3	*X*4	*X*5	*X*6
Mean	161.5677778	86.17481481	49.27444444	79.95370370	35.62222222	25.92296296
StD	3.3015245	1.72576487	2.67475856	1.57855861	0.70326346	0.63925326

由表 5-1 至表 5-4 的数据可以看出，第一主成分贡献率达到 69.11%，前两个主成分累计贡献率已经达到 84.18%，所以前两个主成分已经可以表现生长发育的大部分信息。虽然这样做会损失一部分信息，但是由于它使分析者抓住了主要矛盾，并从原始数据中进一步提取了某些新的信息，因而在某些实际问题的研究中得益比损失大，这种既减少了变量的数量，又抓住了主要因素的做法有利于问题的分析和处理。

表 5-3　相关矩阵的特征值

项　目	Eigenvalue	Difference	Proportion	Cumulative
1	4.14684049	3.24283040	0.6911	0.6911
2	0.90401009	0.55804122	0.1507	0.8418
3	0.34596887	0.08095599	0.0577	0.8995
4	0.26501288	0.05770141	0.0442	0.9436
5	0.20731147	0.07645526	0.0346	0.9782
6	0.13085621		0.0218	1.0000

表 5-4　相关矩阵的特征向量

项　目	*Z*1	*Z*2	*Z*3	*Z*4	*Z*5	*Z*6
X1	0.405262	-.457808	0.173575	0.066498	0.727366	0.250124
X2	0.432479	-.322250	0.067872	0.184153	-.648690	0.499799
X3	0.438029	-.168803	0.166924	-.714428	-.165785	-.462473
X4	0.378699	0.522103	-.562490	-.308244	0.146324	0.388813
X5	0.443444	-.24585	-.422174	0.550118	-.017081	-.567098
X6	0.341759	0.620383	0.665414	0.231548	0.030873	-.031335

资料来源：惠晓萍. SAS 软件在医学统计主成分分析中的应用. 中国现代医药杂志 2011 年 8 月第 13 卷第 8 期.

主成分的表达式如下：

$$Z1=0.405262X1+0.432479X2+0.438029X3+0.378699X4+ 0.443444X5+0.341759X6$$

$$Z2=-0.457808X1-0.322250X2-0.168803X3+0.522103X4-0.024585X5+0.620383X6$$

第一主成分的计算系数都是正数，所以它是生长发育各因素的一个加权平均，可以解释为魁梧程度，身高、坐高、体重、胸围、肩宽、骨盆宽各部分数值大魁梧程度就大。由第二主成分可以看出胸围和骨盆宽的系数为正，其余指标的系数为负，这也说明能主要从胸围和骨盆宽的指标来评价男孩的发育情况。

SAS 医学科研统计分析软件包是一个组合式软件包。它集数据整理、分析过程、结果输出等功能于一体，使用程序方式，用户可以完成所有需要做的工作，包括统计分析、预测、建模和模拟抽样等。它使用 Windows 窗口方式展现各种管理和分析数据的功能，医学科研工作者只要掌握相应的 Windows 操作技能，粗通医学统计分析原理，就可以得到分析后的数据结果，还可以得到直观、清晰的统计图，形象地展示了对原始数据和分析结果的各种描述，并得出结论。

5.2 Weka 软件介绍

（1）Weka 软件

Weka 的全名是怀卡托智能分析环境（Waikato Environment for Knowledge Analysis），是一款免费的、非商业化的（与之对应的是 SPSS 公司商业数据挖掘产品 Clementine ）、基于 JAVA 环境下开源的机器学习（machine learning）以及数据挖掘（data minining）软件。它和它的源代码可在其官方网站下载。有趣的是，该软件的缩写 WEKA 也是 New Zealand 独有的一种鸟名，而 Weka 的主要开发者同时恰好来自 New Zealand 的 the University of Waikato。软件如图 5-2 所示。

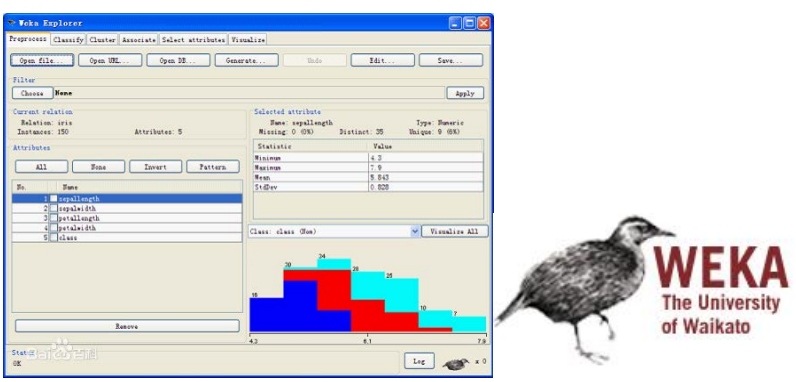


图 5-2　Weka 软件

（2）运用案例

分类技术在各种自动化中很受欢迎。在早期阶段，肝硬化患者的问题是不容易发现，因为即使

它是部分受损，它也会运作正常。肝脏问题的早期诊断能提高患者的生存率。通过分析血液中酶的水平，可诊断肝脏疾病。Michael J Sorich 指出，SVM 分类产生最佳预测的性能化学数据集。黄隆诚认为朴素贝叶斯分类器比 SVM 和用于疾病预防控制中心的慢性疲劳综合征集的 C4.5 算法产生更高性能。保罗哈拍认为没有必要有一个单一的最好的分类工具，而表现最好的算法将取决于要分析的数据集的功能。

分类技术是一种根据输入数据集建立分类模型的系统方法。分类法的例子包括决策树分类法、基于规则的分类法、神经网络、支持向量机和朴素贝叶斯分类法。本文使用决策树分类法进行分类。首先在给定的训练集中，基于 C4.5 算法确定分类模型，该模型能够很好地拟合输入数据，再通过测试集来检测该模型分类的正确性。若拟合较好，便可以用来对未知的分类进行预测。

C4.5 算法是构造决策树分类器的一种算法，能够处理描述性属性是连续型的情况。这种算法利用比较各个属性的 Gain 值的大小，来选择 Gain 值最大的属性进行分类。如果存在连续型的描述性属性，那么首先要做的是把这些连续性属型的值分成不同的区间，即“离散化”。

C4.5 算法继承了 ID3 算法的优点，并在以下几方面对 ID3 算法进行了改进：

① 用信息增益率来选择属性，克服了用信息增益选择属性时偏向选择取值多的属性的不足；

② 在树构造过程中进行剪枝；

③ 能够完成对连续属性的离散化处理；

④ 能够对不完整数据进行处理。

（3）Weka 操作步骤

数据集说明：本文数据来源于加州大学欧文机器学习库的 ILPD（Indian Liver Patient Dataset），数据包含 583 条记录，11 个属性。该数据集包含 416 个肝病患者的病历和 167 个非肝病患者的记录。数据集从印度安得拉邦的东北部收集。该数据集包含 441 男性病人的病历和 142 名女病人的病历。随机选择其中 100 条作为训练集，其余 483 条作为测试集。预处理具体步骤如下。

第一步：打开训练集后，先对要分类的属性进行处理。由于 group 是 numeric 型，C4.5 等算法不能处理数值型。所以把它变为 nominal 型，在 filter 里选择 Numeric To Nominal 其次要对连续的属性离散化，在 filter 里选择 Discretizeo。

第二步：classify，选择 using training set，J48（即为 C4.5），从混淆矩阵可以看出，训练集 100 个样本中有 77 个肝病患者，和 23 个非肝病患者。有 3 个肝病患者预测错误，1 个肝病患者预测错误。

不同分类方法的检测结果如表 5-5 所示。

表 5-5　三类分类方法的校验结果比较

	决策树	*K* 最近邻	朴素贝叶斯
效验准确率	96%	68%	67%
训练	a　b　←classified as	a　b　←classified as	a　b　←classified as
混淆矩阵	74　3	a=1	62 15	a=1	58 19	a=1
	1 22	b=2	17　6	b=2	14　9	b=2

资料来源：黄速成. 基于 Weka 平台的决策树在医学中的应用. 北京物资学院.

结论：决策树分类效果较好。

可以用以上分类规则对患者进行判别分析，如：根节点是 A/G Ratio Albumin and Globulin Ratio，在(0.91-0.975]区间的属于非肝病患者，有两个样本。A/G Ratio Albumin and Globulin Ratio 不在此区间的，若 TP Total Protiens 在(5.25-5.65]且 Sgpt Alamine Aminotransferase 在区间(109-226]的属于第一类 1，有 3 个样本，若 TP Total Protiens 在(5.25-5.65]且 Sgpt Alamine Aminotransferase 不在(109-226]的属第二类 2，有 7 个样本，置信度为 1.0。

5.3 Matlab 案例

（1）Matlab 软件

Matlab 是美国 MathWorks 公司出品的商业数学软件，用于算法开发、数据可视化、数据分析以及数值计算的高级技术计算语言和交互式环境，主要包括 Matlab 和 Simulink 两大部分。

Matlab 是 matrix&laboratory 两个词的组合，意为矩阵工厂（矩阵实验室）。是由美国 mathworks 公司发布的主要面对科学计算、可视化以及交互式程序设计的高科技计算环境。它将数值分析、矩阵计算、科学数据可视化以及非线性动态系统的建模和仿真等诸多强大功能集成在一个易于使用的视窗环境中，为科学研究、工程设计以及必须进行有效数值计算的众多科学领域提供了一种全面的解决方案，并在很大程度上摆脱了传统非交互式程序设计语言（如 C、FORTRAN）的编辑模式，代表了当今国际科学计算软件的先进水平。

Matlab 和 Mathematica、Maple 并称为三大数学软件。它在数学类科技应用软件中在数值计算方面首屈一指。Matlab 可以进行矩阵运算、绘制函数和数据、实现算法、创建用户界面、连接其他编程语言的程序等，主要应用于工程计算、控制设计、信号处理与通讯、图像处理、信号检测、金融建模设计与分析等领域。

Matlab 的基本数据单位是矩阵，它的指令表达式与数学、工程中常用的形式十分相似，故用 Matlab 来解算问题要比用 C、Fortran 等语言完成相同的事情简捷得多，并且 Matlab 也吸收了像 Maple 等软件的优点，使 Matlab 成为一个强大的数学软件。在新的版本中也加入了对 C、Fortran、C++、Java 的支持。软件如图 5-3 所示。

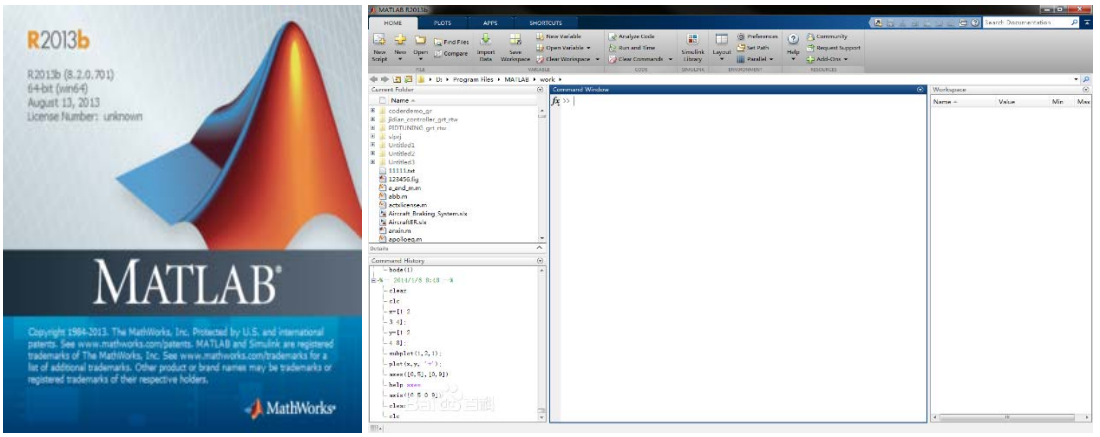


图 5-3　Matlab 软件

（2）案例

背景：心肌梗塞（MI）是心血管疾病之一，其发病机理如下。是发生在冠状动脉粥样硬化病变的基础之上的，由此病症并发出出血、粥样斑块破裂、血管腔内血栓形成等病症，动脉内膜下的出血和动脉持续痉挛都是导致管腔阻塞的原因，当管腔发生阻塞时，如果该病变的动脉没有其他的支循环连接的话，该动脉所供应的心肌就会产生持久而且十分严重的缺血，当缺血时间达到了一个小时以上的话将会导致此处的心肌完全坏死，人的身体健康和生命安全都会有危险。发生心肌缺血有以下的情况，在已患有粥样硬化的冠状动脉又发生了管腔狭窄，如此便会使得心脏排血量骤降，或者左心室所承担的负荷量剧增（原因可能有做过于繁重的体力活动、用力大便或血压剧升）时，也可使心肌产生严重缺血，持续的时间长久的话将引起心肌严重坏死。当人食入大量脂肪类食物后使得体内的血脂量增加，血液的粘稠性也增高，血液的流动性减小，使得血液流动缓慢，血小板易于聚集而致血栓形成；另外在人进入睡眠状态时，神经的紧张程度增高，冠状动脉发生痉挛，都会导致心肌坏死。心肌梗塞发生的群体也是不确定的，可能是发生在从未患过此病的人身上，也可能发生在有过心梗的病人身上。

方法： ST-T 段准确的截取对于早期心肌梗塞诊断十分重要，ST-T 段在临床中广泛用于早期心梗的检测。在千奇百怪的心电数据的 ST-T 段截取中，T 波的方向可能是正向的，也有可能是负向的，为了让截取更加准确，本文在 T 波初始点检测中运用到了面积法，对于 T 波方向可以不用考虑，能够更加准确地截取 ST-T 段。如图 5-4 是通过自动截取 ST-T 得到的截取结果，其中蓝色部分是本文自动截取出来的 ST-T 段数据。

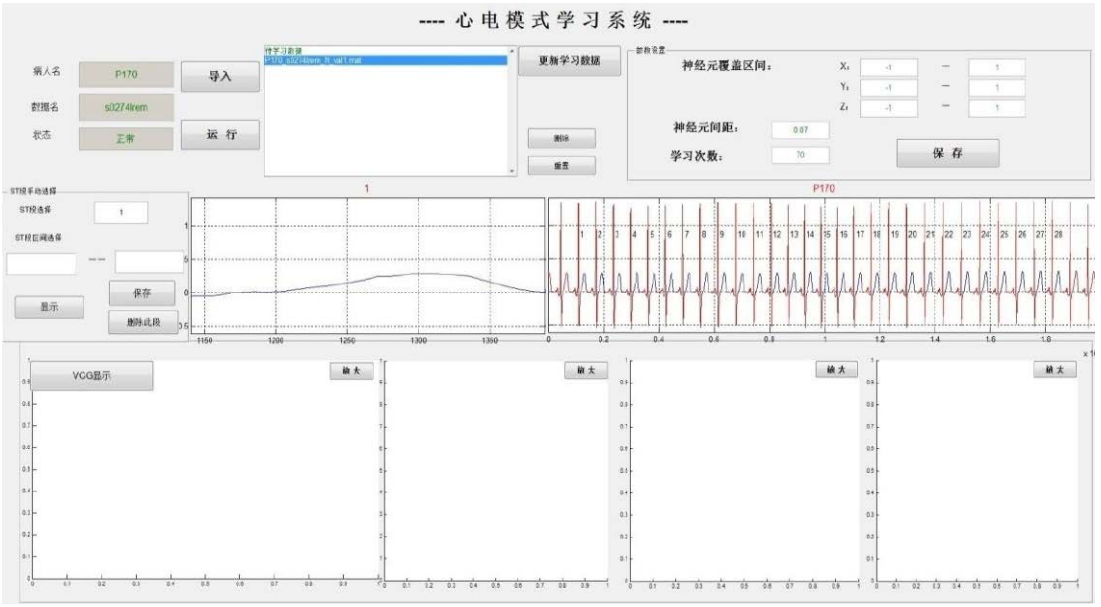


图 5-4　心电模式学习系统界面

资料来源：万志云. 早期心肌梗塞辅助诊断系统模式学习及模式库管理的 Maltab 实现. 华南理工大学. 2013 年华南理工大学硕士论文.

对单个 ST-T 段进行修改首先需要确定修改的 ST-T 段的位数为多少，然后在 ST 段选择栏中输入需要修改的段值，然后单击“显示”按钮可以查看该 ST-T 段的放大效果，在 ST 段区间选择栏中，左边是起始点即 J 点，右边是终止点即 K 点，当选择栏中输入有指定的点值再单击“显示”按钮，则显示指定的点的位置，若其中为空，则显示自动截取中的位置，其中手动调整有一个原则是左、右空格栏中的值不能超过此段的终止点、起始点的范围，否则会提示错误。用以上的原则来调整好指定的 ST 段起止点后，单击“保存”按钮，就完成了该 ST 段的手动调整的工作，如果起止栏中有空的（即没有修改该 ST 段的起点或者终点），则不修改相应点的位置。

如图 5-5 和图 5-6 所示，我们介绍了单个心电数据导入学习建模的过程，同时也提到了心电模式学习系统具有批量学习的功能。在平常的建模过程中时常需要一次性导入几个数据，或者一次性导入一种类型的数据进行建模，这样一个个的点击导入很浪费时间，由此我们引入了 SQL Serve：与 Matlab 连接，将所有采集的病人信息进行统计归纳到数据库中，再通过 Matlab 的界面设计来完成批量的模式导入工作，从而来减轻工作量。

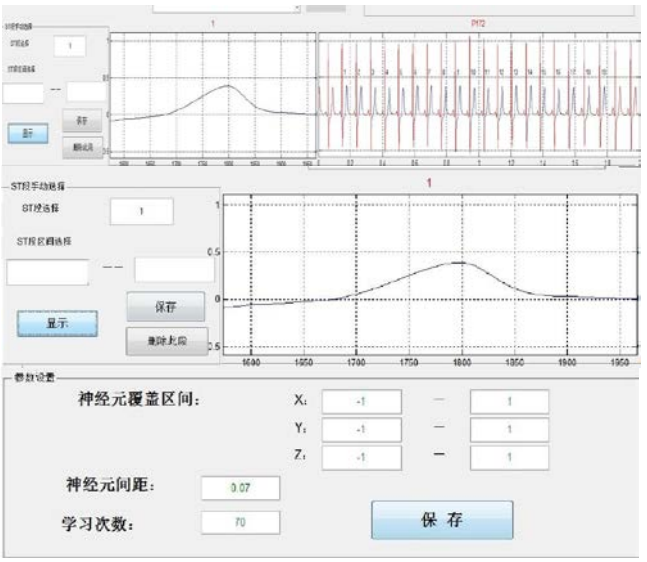


图 5-5　学习效果图

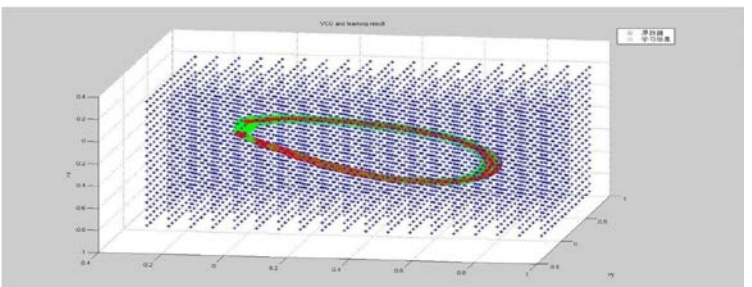


图 5-6　学习效果图

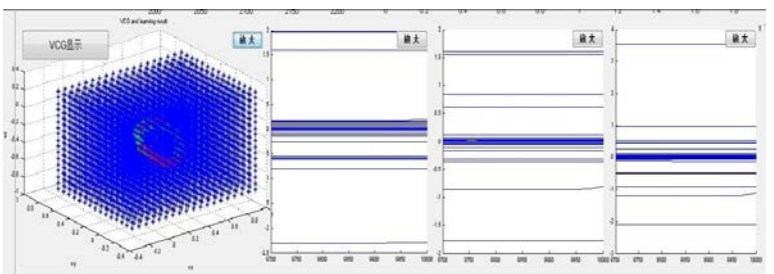


图 5-7　VCG 显示图界面

Matlab 与 SQL Server 通讯能够采用 JDBC 和 ODBC 这两种方法，但是由于本系统中使用的 Matlab 是 64 位的，64 位的 Maltab 软件与 SQL Server 通过 ODBC 方式连接第一是速度慢，第二是 Matlab

没法通过 ODBC 接口从 SQL Server 中提取数据，所以本文采用的是 JDBC 的连接方式与 SQL Server 进行访问。通过网站下载相应的驱动，然后配置相应的路径，就可以成功实现连接。经过 SQL Server 去数据库中读取数据完成三步便可。

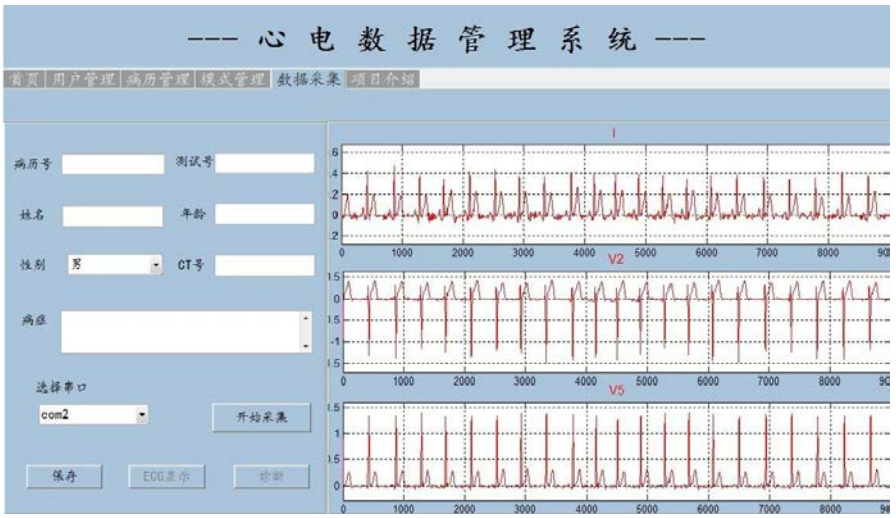


图 5-8　数据采集界面

资料来源：万志云. 早期心肌梗塞辅助诊断系统模式学习及模式库管理的 Maltab 实现华南理工大学. 2013 年华南理工大学硕士论文.

如图 5-7 和图 5-8 所示，本系统不仅能对一个心电数据进行学习建模，为了方便快速建立模式库，本系统可对批量的心电数据进行学习、建模，只需要将这些批量学习的数据做好数据预处理操作即可。如图 5-9 所示，在显示框中可以看到即将要学习的数据，同时在显示框旁边有“删除”和“重置”按钮，当需要删除某一个已加入学习的数据时，只需选中要删除的数据，单击“删除”按钮即可进行删除此即将学习数据。重置按钮实现的是将整个导入显示框中要学习的数据全部清空，重新进行添加需要学习的数据。对每个数据的 ST-T 段截取的修改可以在逐个导入时修改，也可以在全部导入后再返回去对其中任意一个数据进行修改，使得系统更加智能。在学习中，图 5-10 的左侧显示的是正在学习的数据，可以看到学习进度。

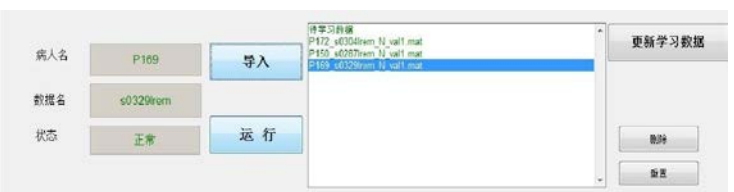


图 5-9　数据导入界面

图 5-10　生成学习表

资料来源：万志云. 早期心肌梗塞辅助诊断系统模式学习及模式库管理的 Maltab 实现. 华南理工大学. 2013 年华南理工大学硕士论文.

（3）计算机操作分解

① 数据采集

a. 采集模块：选定好心电采集盒连接的串口号后，单击“开始采集”按钮便可进行数据的采集，如果串口选择错误，系统会提示打开串口失败。此操作的流程是串口的打开、串口参数设置、发送命令（开始、停止）、获取缓冲区数据并转换、关闭串口。

b. 打开串口

创建串口对象：scorn = serial（com_num）；其中 corn num 是串口号。设置好后，在 Matlab 中打开串口的指令是 fopen（scom）。

c. 串口参数设置：波特率为 115 200bps，1 为停止位，8 位数据位，无校验。

用如下语句进行此串口参数设置。

set（scorn，'BaudRate'，baudrate，'Parity'，parity，'DataBits'，databits，...'StopBits'，stopbits，'BytesAvailableFcnCount'，10，…'BytesAvailableFcnMode'，'byte'，'BytesAvailableFcn'，{@ bytes，

handles}）;

d. 发送开始命令：上位机向心电模块发送的命令总共有 20 字节，在这 20 个字节中不要的附加字节需要用 0 填充，其命令的格式如下：

AIKD COx41, 0x49, Ox4b, 0x44)+命令(1)+附加字节(14)+校验(1)

(0) (1) (2) (3) (4) (5-18) (19)

以上的命令为 16 进制的方式，根据附加字节值不同对心电采集盒的控制命令不同，本文中用到的开始、停止命令如下，用十进制的方式存储的矩阵数组。

开始（start）：65737568211000000000000047

停止（stop）：65737568210000000000000046

在采集过程中会有一个采集进度条，来提示采集的进程，当进度条走完即采集完成，程序上最后用 fwrite（scorn，start，'uint8'，'async'）这个语句来将控制命令发送到心电采集盒，由此来对心电采集盒进行控制。

此采集盒的命令中除了开始与停止命令外还有许多命令，起搏检测导联命令、复位命令、握手命令等都是可以通过上位机来对采集盒控制的。同时，采集盒本身也有硬件滤波的功能，可以通过发送相应的命令对采集数据进行工频滤波、基漂滤波、肌电滤波，从而来得到准确、干扰小的心电数据。

e. 获取缓冲区数据并转换：模块每 2ms 给上位机发送一个数据包，每个数据包有 16 字节。我们需要通过对这 20 秒的数据进行数据的转换，做相应的数据操作才能得到标准的心电数据。同时，由于一次将 35 秒的心电数据从缓冲区中取出再进行一次性的转换速度很慢，数据量太大了，所以本文选择对数据分块进行处理来加快数据转换的速度。

f. 关闭串口：用 fclose （scorns）语句将串口关闭。

病人信息录入模块：在图 5-8 中可以看到，此模块是为了记录病人数据的相关信息，包括病历号、测试号、姓名、年龄、性别、CT 号、病症，其中病历号和测试号这两栏为必填项。当采集完数据，同时输入完病人信息后，单击“保存”按钮可以将采集的数据保存到数据库中，然后可以对此数据做进一步的操作。可将此数据导入到识别子系统中，来对此数据进行诊断。当单击“诊断”按钮进入识别系统中后，在识别界面中单击“更新采集数据”按钮便可对此采集数据进行诊断，如此加快数据诊断的时间，避免了离线的数据转换工作，使得系统整体化。

② 采集结果显示模块：为了便于查看采集结果，在数据采集完成后，系统界面显示了 ECG 的 I, V2, VS 这三个导联的采集结果。还可以单击界面中的“ECG 显示”按钮来查看采集的 12 个导联心电图，如图 5-11 所示。ECG 的显示方式是按照传统的临床心电图纸格式进行设计的，取了 35 秒数据中的几个周期来进行显示，一般取其中的 10 个左右的心跳周期。

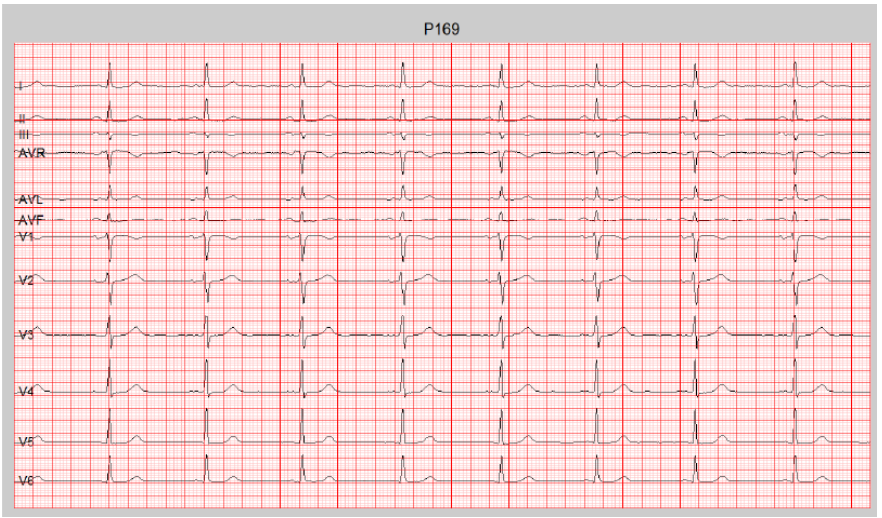


图 5-11　ECG 显示

③ 心电数据识别诊断设计：我们介绍了此系统的学习过程，学习过程与识别过程心电数据预处理的操作都是相同的，即都需做 ECG 转 VCG、滤波及 ST 段截取，在本章中的系统识别部分预处理内容就不再赘述。识别过程是此系统的核心部分，基于心电图 ST-T 段的早期心肌梗塞辅助系统的识别过程实际上就是从大量的模式库找出与被测量的导入数据最为相似的模式库中数据，该被测数据的病症与识别出来的模式库中数据的病症最为相似，由此来得到该被测数据的诊断结果，下面将具体描述识别过程的实现。

通过动态模式识别方法，根据被测数据中提取的特征与模式库中的数据进行匹配，从而得到残差最小的前九个匹配，也就是我们所说的匹配结果了，在识别界面中做好了预处理操作后单击“运行”按钮，系统就开始进行诊断。实例 T057 的诊断匹配结果如图 5-12 所示，从图中可以看到“第一匹配”选项的内容是空的，是因为本系统将匹配到本身的模式库数据进行了屏蔽不显示。我们可以看到 T057 的其他八个匹配中都有相应匹配数据的大致情况，如第三匹配 P260 就是正常的数据。

图 5-12　匹配结果

在识别结果中还可查看残差图，来看此数据与模式库中数据的残差值为多少，实例 T057 的残差结果如图 5-13 所示，在图中的右上角有标出第一到第五匹配的值在残差图中的位置。

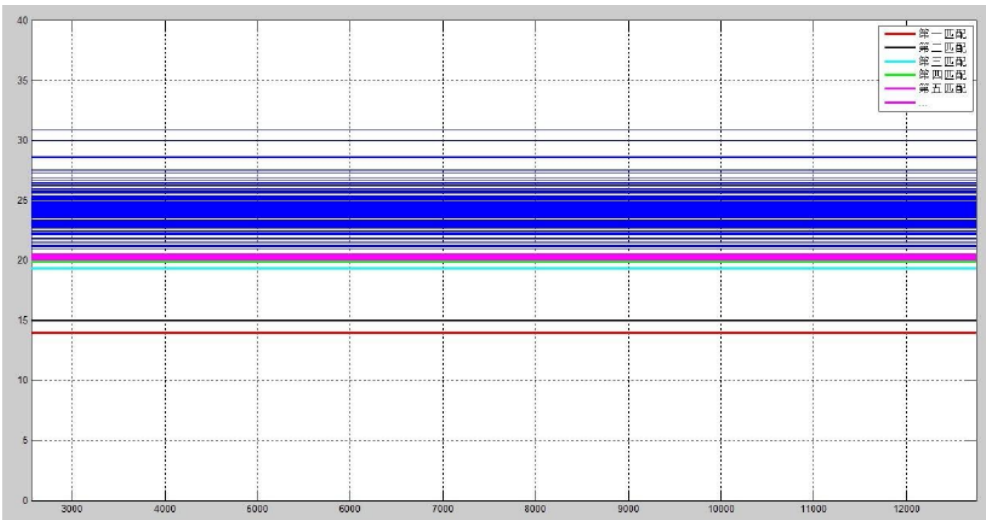


图 5-13　残差图

资料来源：万志云. 早期心肌梗塞辅助诊断系统模式学习及模式库管理的 Maltab 实现. 华南理工大学. 2013 年华南理工大学硕士论文

通过查看这些整体的匹配结果还不能给出一个明确的诊断结果，其原因有两个：第一，识别出来的数据有些并不是与被测数据在各方面都类似的，这些匹配出来的结果需要被删除；第二，由于我们现在建立的模式库数量还有限，没有涵盖所有可能的心电数据模型，导致有些被测数据不能找到与其最匹配的数据。基于这两点我们有在做相应的工作，对于第一点我们下面会有讲到对匹配结果的筛选，针对第二点我们现在正在不断地采集临床数据，用这些数据来丰富心电模式库。

由于模式库中的数据量有限，该系统诊断识别出来的部分数据可能不是相互最匹配的数据，所以我们需要对诊断结果进行筛选，我们需要单击不同匹配的按钮来查看被测数据与该匹配结果更多的对比信息，例如我们查看 T057 识别出来结果中的第六匹配，单击匹配结果栏中的“第六匹配”按钮，就可以查看 T057 与 T224 这两个数据的各项信息的对比，图 5-14 即是对匹配结果进行筛选的界面。通过三方面进行判断，即两数据的 ECG 各个导联的对比、两数据 ST-T 环进行对比同时与匹配出来的 9 个匹配结果整体比较，此两项来对该匹配数据进行筛选，同时还可以对这个匹配结果进行备注。

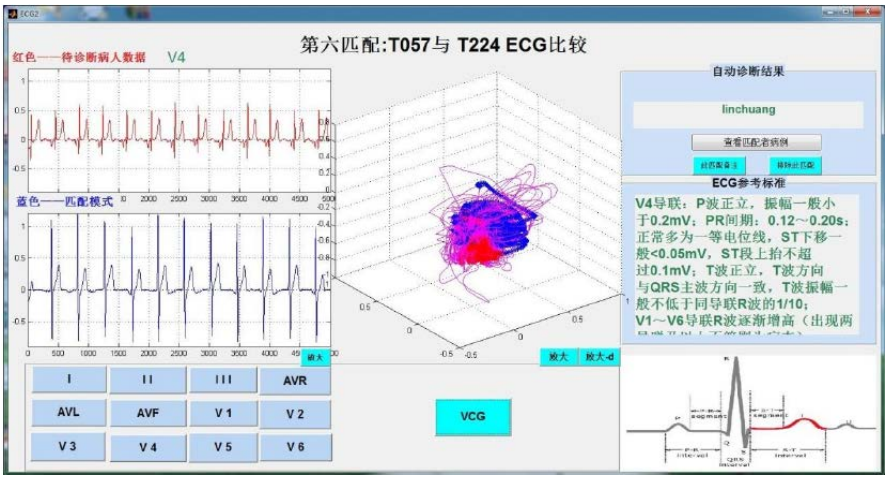


图 5-14　匹配结果筛选界面

资料来源：万志云. 早期心肌梗塞辅助诊断系统模式学习及模式库管理的 Maltab 实现. 华南理工大学. 2013 年华南理工大学硕士论文.

通过匹配结果的筛选操作以后可以得到更加准确的匹配结果，来提高医生的辅助判断的效果，当做完匹配结果筛选以后，可以得到被测数据的初步诊断结果，将得到该被测数据的心电图、心向量图、初步诊断的文字描述信息，填到图 5-8 早期心肌梗塞诊断系统界面的左下角的病人信息模块中，填写完后可以将此被测数据的病例报告进行打印，打印报告中的内容与病人信息输入模块中的信息一样，打印成纸质版便于存档管理，同时可以添加到数据库中。由此完成了采集数据的诊断部分，整个过程都只是在对心电数据做处理，此种早期心肌梗塞辅助诊断是一种无创的检查，这避免了病人高额的医疗费用，具有较高的医学意义。

这是一个典型的利用计算机实现智能电生理信号数据挖掘的案例，以 Matlab 来实现人工智能的诊断。其原理也很简单，通过对脉冲信号的分段截取，实现数据的聚类从而用编程语言 Matlab 来实现标准对照库的建立。建立数据模型（分型）的对照库后，用匹配的方法来对分型进行模式识别，这是一种常用的工程实现手法。值得注意的是，目前的机器学习工具在模糊理论，不确定性概率算法上与人脑相比还有很大的差距，人脑可以根据简单的推理，碎片抽象与组合的强大能力来还原一个个事实真相，机器在面对不确定性、模糊理论的时候，特别是在心电波信号的模式识别上，人脑的经验累积有很强的不连续性特点。有时，一个看似正常的电波信号截断，机器可以模式化地批处理进入正常分型，而有经验的医生可以从细微的变化中发现蛛丝马迹从而产生异常的疑问，因为这后面暗藏着他多年的脑海中一个个鲜活的面孔与病例，这些是机器短期内无法拟合与逼近的。因此，用数据的模式识别来代替心梗的无创检查还有很远的路要走。

5.4 R 语言案例

R 语言是一个用于统计计算及统计制图的优秀的开源软件，也是一个可以从大数据中获取有用信息的绝佳工具。它能在目前各种主流操作系统上安装使用，并且提供了很多数据管理、统计和绘图函数。

下面本节就将使用 R 语言所提供的强大的函数库来构建一棵决策树并加以剪枝。

清单 1. 构建决策树及其剪枝的 R 代码

```
library("rpart")library("rpart.plot")library("survival")# 查看本次构建决策树所用的数据源
stagec# 通过 rpart 函数构建决策树
fit <- rpart(Surv(pgtime, pgstat)~age+eet+g2+grade+gleason+ploidy,stagec, method="exp")# 查看决策树的具体信息
print(fit)printcp(fit)# 绘制构建完的决策树图
plot(fit,uniform=T, branch=0.6, compress=T)text(fit, use.n=T)# 通过 prune 函数剪枝
fit2 <- prune(fit, cp=0.016)# 绘制剪枝完后的决策树图
plot(fit2, uniform=T, branch=0.6, compress=T)text(fit2, use.n=T)
```

根据代码，运行步骤如下：

① 导入需要的函数库。当然如果本地开发环境没有相应的库的话，还需要通过 install.packages 函数对库进行安装。查看本次构建决策树的数据源。stagec 是一组前列腺癌复发的研究数据。通过 rpart 函数构建决策树，以研究癌复发与病人年龄、肿瘤等级、癌细胞比例、癌细胞分裂状况等之间的关系。查看决策树的具体信息。绘制构建完成的决策树图。通过 prune 函数对该决策树进行适当的剪枝，防止过拟合，使得树能够较好地反映数据内在的规律并在实际应用中有意义。

② 绘制剪枝完后的决策树图。

该案例决策树的拟合结果如下，剪枝前后的树如图 5-15 和图 5-16 所示。

```
n=  146
node), split, n, deviance, yval
      *denotes terminal node
1) root 146 192.111100 1.0000000
  2) gr ade<2.5 61  44.799010 0.3634439
    4) g2<11.36 33  9.117405 0.1229835*
    5)g2>=11.36 28  27.602190 0.7345610
     10)gleason<5.5 20 14.297110 0.5304115*
     11)gleason>=5.5 8  11.094650 1.3069940*
  3)grade>=2.5 85  122.441500 1.6148600
   6)age>=56.5 75  103.062900 1.4255040
    12)gleason<7.5 50  66.119800 1.1407320
      24)g2<13.475 24  27.197170 0.8007306*
```

```
        25)g2>=13.475 26  36.790960 1.4570210
          50)g2>=17.915 15  20.332740 0.9789825*
          51)g2<17.915 11  13.459010 2.1714480*
       13)gleason>=7.5 25  33.487250 2.0307290
         26)g2>15.29 10  11.588480 1.2156230*
         27)g2<15.29 15  18.939150 2.7053610*
      7)age<56.5 10  13.769010 3.1822320*
```

如图 5-15 所示，决策树为什么（WHY）要剪枝？原因是避免决策树过拟合（Overfitting）样本。前面的算法生成的决策树非常详细并且庞大，每个属性都被详细地加以考虑，决策树的树叶节点所覆盖的训练样本都是“纯”的。因此用这个决策树来对训练样本进行分类的话，你会发现对于训练样本而言，这个树表现完好，误差率极低且能够正确地对训练样本集中的样本进行分类。训练样本中的错误数据也会被决策树学习，成为决策树的部分，但是对于测试数据的表现就没有想象得那么好，或者极差，这就是所谓的过拟合（Overfitting）问题。Quinlan 教授试验，在数据集中，过拟合的决策树的错误率比经过简化的决策树的错误率要高，见图 5-16。

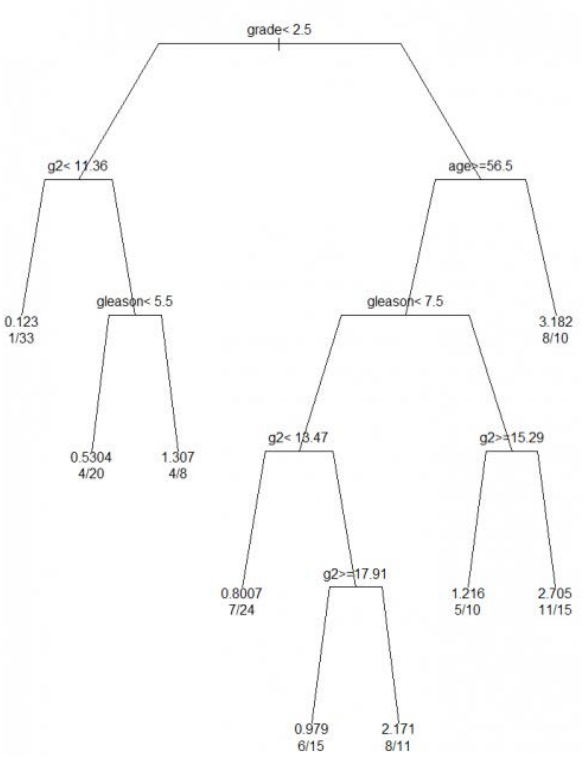


图 5-15　未剪枝的决策树

资料来源：来自 36 大数据（36dsj.com)

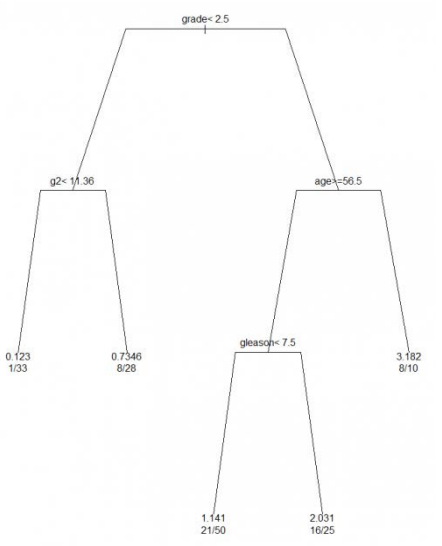


图 5-16　剪枝后的决策树

资料来源：来自 36 大数据（36dsj.com)

R 语言简单易学，比如语言结构相对松散，使用变量前不需明确正式定义变量类型，但仍保留了程序设计语言的基础逻辑与自然的语言风格，最重要的是开源，在社区容易获得。R 语言十分适合在校医学院学生的学习使用，对于临床医生而言，SPSS、SAS、Matlab、Weka 这些更为实用。

5.5 临床医生如何用好挖掘工具

虽然在现代的医学教育范式下医学统计学已经成为各大医学院校的必修课，高年资医生的医学统计学修养已经有一定的高度，但我们仍然在实践中发现许多优秀的临床医生懂一点统计学却不懂数据挖掘，喜欢倚重自身的临床经验而没有数据，喜好下结论却写不好论文，这些都是临床医生不懂得使用挖掘工具的结果。

下面用一个故事来看看临床医生如何使用好挖掘工具。

某三甲医院科研处处长向医院院长报送一份近期的医院用药情况报告如下：

近来，随着低出生体重儿的存活率提高及抗生素的广泛应用，机会菌感染引起的新生儿败血症有所增加。现对我院新生儿病房收治 14 例资料完整患儿报告如下，以提高治愈率。

临床资料

我院新生儿病房近 4 年共收治新生儿败血症经血培养确诊 48 例，其中血培养为表皮葡萄球菌 14 例……

应当说，这样的报告很平庸，不仅数据案例远远不足，而且由于没有应用数据库技术和挖掘工具，其背后的数据处理理念十分落后，依然停留在统计总结的传统时代。事实上，今天的三甲医院有很好的信息化基础，完全可以使用现代的数据挖掘技术来依据本院的用药情况完成一份科学的报告。我们来看一个完整的案例。

案例：基于R语言的基层门诊用药大数据分析

目的：利用卫生信息技术探索基层医疗卫生机构门诊用药规律，为卫生管理决策和临床合理用药提供依据。

方法:通过四川省基层医疗机构管理信息系统,提取某县5 个乡镇卫生院门诊2012 年9 月～2014 年3 月用药，采用R 语言对门诊大数据进行分析。结果：基层门诊使用最多的是维生素 B6 片、维生素 C 片、头孢克肟分散片等类药物，药品合并使用情况明显。

结论：利用信息技术手段，通过统一的基层医疗机构管理信息系统，可加强卫生行政部门对基层用药监管，促进基层合理用药。

资料数据搜集：使用四川省基层医疗机构管理信息系统，通过 SQL 语句从基层系统提取四川省某县 5 个乡镇卫生院 2012 年 9 月 1 日～2014 年 3 月 1 日的门诊记录 100 505 条。提取信息为就诊日期、就诊者性别、年龄、诊断、门诊用药等。

使用 R语言 3.1.0 版本，配置 arules 关联规则分析包，eclat 函数计算频繁项，apriori 函数挖掘关联规则，参数设置 s（支持度）为 0.03、c（置信度）为 0.02。数据格式为“购物篮”格式，将单次门诊处方各药品进行分析，观察不同药品的“共现”情况，分析数据的折半频繁项和关联规则。

严格纳入标准，排除单次门诊重复用药、门诊未用药的数据和该县某一儿童专科乡镇卫生院（因该院绝大多数患者为儿童，不能代表国内一般性乡镇卫生院的用药情况）。纳入数据量相对较大、时间跨度长，样本具有一定的代表性。

结果与分析：100 505 条门诊用药记录中，纳入就诊者中男性比例（58.85%）较女性（41.12%）偏高。以 40～60 岁的中年人最多，其次是＜10 岁和 60～80 岁年龄段就诊者。门诊就诊者年龄构成如表 5-6 所示。随着四川省基层医疗机构管理信息系统逐步推广，纳入门诊量呈逐步递增趋势，其中 2013 年 12 月份纳入门诊用药量最大。通过门诊用药数据计算出门诊均次用药品种数为 5. 022 种，单次门诊用药数量最多为 18 种药品，大于 10 种药品的占总门诊的 0.43%。

表 5-6　门诊就诊者年龄构成

年龄段	门诊数	比例/%
<10	11 575	11.52
10-	4 328	4.31
20-	6 457	6.42
30-	8 019	7.98
40-	18 804	18.71
50-	15 788	15.71

续表

年龄段	门诊数	比例/%
60-	17 267	17.18
70-	11 171	11.11
≥80	4 512	4.49
未知	2 584	2.57
合计	100 505	100

频繁项分析：单药频繁项通过 arules 包中的 eclat 函数，设置参数最小（支持度 S 为 0.04）求频繁项集，即在门诊用药记录中，某种药品出现频率，按照由大到小的次数排序，得到门诊使用较多（频繁）的药品列表（表 5-7）所示。从表 5-8 可发现门诊用药以片剂、胶囊口服剂型为主，维生素 B6 片（24.24%）、维生素 C 片（22.38%）、头孢克肟分散片（19.61%）使用最为频繁，门诊处方超过 10%的有 11 种药品。

表 5-7　基层门诊开具的单药频繁项

序　号	药　物	支持度
1	维生素 B6 片	0.2424
2	维生素 C 片	0.2238
3	头孢克肟分散片	0.1961
4	盐酸澳已新片	0.1577
5	对乙酰氨基酚片	0.1516
6	奥美拉唑肠溶胶囊	0.1498
7	盐酸异丙嗪片	0.1217
8	多潘立酮片	0.1140
9	马来酸氯苯那敏片	0.1056
10	维生素 B1 片	0.1018
11	醋酸泼尼松片	0.1000
12	西咪替丁片	0.0990
13	感冒清片	0.0936
14	布洛芬缓释胶囊	0.0898
15	氨茶碱片	0.0890
16	消旋山莨若碱片	0.0884
17	头孢呋辛脂胶囊	0.0807
18	头孢拉定胶囊	0.0797
19	醋酸地塞米松片	0.0752
20	咳特灵胶囊	0.0690

资料来源：王帅，林晓东等“基于R语言的基层门诊用药大数据分析”中华医学图书情报杂志 2015 年 3 月第 24 卷第 3 期

多药频繁项：同样通过 arules 包中的 eclat 函数，设置参数最小（支持度 S 为 0.04）求多种药物（2 种）的频繁项集，即在门诊用药记录中，某 2 种药品合并使用出现频率，按由大到小的次数排序，得到门诊 2 种药物合并频繁项集，如表 5-8 所示。从表中可看出，头孢克肟分散片与维生素 C 片合并使用比例约为 6.99%，奥美拉唑肠溶胶囊与多潘立酮片合并使用比例约为 6.38%，合并使用比例超过 5%的有 6 种药品组合，维生素 C 和维生素 B6 与其他药品联合使用的比例较高。

表 5-8　基层门诊合并使用 2 种药物的频繁项

序　号	合并使用药物	支持度
1	头孢克肟分散片，维生素 C 片	0.0699
2	奥美拉唑肠溶胶囊，多潘立酮片	0.0638
3	对乙酰氨基酚片，盐酸澳已新片	0.0597
4	维生素 C 片，盐酸澳已新片	0.0576
5	头孢克肟分散片，维生素 B6 片	0.0565
6	盐酸澳已新片，盐酸异丙嗪片	0.0529
7	对乙酰氨基酚片，头孢克肟分散片	0.0463
8	对乙酰氨基酚片，维生素 C 片	0.0450
9	奥美拉唑肠溶胶囊，维生素 B6 片	0.0429
10	奥美拉唑肠溶胶囊，维生素 B6 片	0.0421
11	维生素 B6 片，维生素 C 片	0.0415

资料来源：王帅，林晓东等. 基于R语言的基层门诊用药大数据分析. 中华医学图书情报杂志 2015 年 3 月第 24 卷第 3 期.

如表 5-9 所示结果显示，活血止痛片和布洛芬缓释胶囊联合使用的比率（s）为 3.06%。在研究门诊数据中，处方开出（前导）活血止痛片后，接着开（后继）布洛芬缓释胶囊的概率是 74.76%，比单独使用布洛芬缓释胶囊的比率提升 8.32 倍（即提升度 lift 为 8.32）。用相同的方法，可解释其他规则。在已有关联规则中，将最大频繁项维生素 C 做为固定后继，设置参数提升度 lift＞1.0，用方法 subset（rules，subset = rhs %in%“维生素 C 片” &lift＞1.0）进行规则挖掘，结果见表 5-10。分析发现，头孢克肟分散片、奥美拉唑肠溶胶囊、布洛芬缓释胶囊、西咪替丁片、多潘立酮片及头孢拉定胶囊 6 种药物联合使用的比率＞30%，这些药品（前导）对维生素 C 片（后继）的影响较大（lift＞1.0）。该分析方法同样适用于其他药品的联合使用研究。

表 5-9　基于 C 排序的合并用药关联规则

序号	前导药物（LHS）	后催药物（RHS）	支持度	置信度	提升度（Lift）
1	活血止痛药	布洛芬缓释胶囊	0.0306	0.7476	8.3224
2	多潘立酮片	奥美拉唑肠溶胶囊	0.0638	0.5596	3.7354
3	醋酸地塞米松片	维生素 C 片	0.0357	0.4751	2.1230

续表

序号	前导药物（LHS）	后催药物（RHS）	支持度	置信度	提升度（Lift）
4	盐酸氨溴索片	盐酸异丙嗪片	0.0311	0.4599	3.7787
5	布洛芬缓释胶囊	维生素 B6 片	0.0399	0.4441	1.8318
6	盐酸异丙嗪片	氨茶碱片	0.0393	0.4418	3.6304
7	咳特灵胶囊	氨茶碱片	0.0301	0.4371	4.9088
8	盐酸异丙嗪片	盐酸溴已新片	0.0529	0.4345	2.7551
9	西咪替丁片	头孢克肟分散片	0.0427	0.4331	2.2077
10	奥美拉唑肠溶胶囊	多潘立酮片	0.0638	0.4257	3.7354

表 5-10　维生素 C 合并用药关联规则

序号	前导药物（LHS）	后催药物（RHS）	支持度	置信度	提升度（Lift）
1	头孢克肟分散片	维生素 C 片	0.0699	0.3565	1.5934
2	盐酸溴已新片	维生素 C 片	0.0576	0.3650	1.6312
3	对乙酰氨基酚片	维生素 C 片	0.0450	0.2969	1.3268
4	感冒清片	维生素 C 片	0.0379	0.4047	1.8085
5	醋酸地塞米松片	维生素 C 片	0.0357	0.4751	2.1230
6	马来酸氯苯那敏片	维生素 C 片	0.0320	0.3030	1.3541

结果显示，基层用药以治疗上呼吸道感染、腹泻、咳嗽等常见病、多发病为主，用药品种相对集中，可为卫生管理部门基本药物遴选、评价、招标、配送等提供参考和为基本药物循证医学评价提供指导。结果显示，维生素类药物使用过于频繁，其中维生素 C 和维生素 B6 片使用最为频繁，且一般作为辅助用药出现，提示可能存在维生素类药物过度使用的情况。但维生素 C 并不是“有百利而无一害”的万能安全药物，与其他药物一样，维生素 C 亦有其适应范围和功能。如果过量使用，只能造成药物浪费，甚至引起不良后果。维生素 B6 用量过大可致新生儿产生维生素 B6 依赖综合症。研究还发现，激素类药物和抗生素类药物使用比较偏高。

本案例采用了信息系统的数据采集，共计 10 万余条数据的处理有效地利用了计算机技术。采用 R 语言数据挖掘包中的关联规则算法对基层用药情况进行了 10 万条数据级别的挖掘，发现了基层医生的用药习惯与待改进行为。这个案例充分说明在计算机与大数据时代，医学论文中几十、数百个病例的小数据时代就要成为过去，在数据挖掘工具的导引下，十万条、百万条乃至千万条数据的大数据时代正在来临，这样的技术趋势必将改变临床数据处理的规则。

第 6 章

专业级医学 SCI 论文中的统计工具

- 医学数据中的 *T* 值与 *P* 值故事
- K 线图的故事
- 国际顶级期刊上的数据技术
- SCI 荟萃分析中的统计学工具

6.1 医学数据中的 *T*值与 *P*值故事

T 指的是 *T* 检验，亦称 student t 检验（Student's t test），主要用于样本含量较小（例如 $n<30$），总体标准差 σ 未知的正态分布资料。

P 值（P value）就是当原假设为真时所得到的样本观察结果或更极端结果出现的概率。如果 *P* 值很小，说明原假设情况发生的概率很小，而如果出现了，根据小概率原理，我们就有理由拒绝原假设，*P* 值越小，我们拒绝原假设的理由越充分。总之，P 值越小，表明结果越显著。但是检验的结果究竟是“显著的”、“中度显著的”还是“高度显著的”需要我们自己根据 P 值的大小和实际问题来解决。

P 值方法的思路是先进行一项实验，然后观察实验结果是否符合随机结果的特征。研究人员首先提出一个他们想要推翻的“零假设”（null hypothesis），比如，两组数据没有相关性或两组数据没有显著差别。接下来，他们会故意唱反调，假设零假设是成立的，然后计算实际观察结果与零假设相吻合的概率。这个概率就是 *P* 值。费希尔说，*P* 值越小，研究人员成功证明这个零假设不成立的可能性就越大。20 世纪 20 年代，英国统计学家罗纳德·费希尔（Ronald Fisher）首次采用 *P* 值方法时，并没有打算把它作为决定性的检验方法。他本来只是用 *P* 值作为一种判断数据在传统意义上是否显著的非正式方法，也就是说，用来判断数据证据是否值得进行深入研究。尽管对 *P* 值提出批评的大有人在，但统计方法的变革仍然进展缓慢。“费希尔、内曼和皮尔森提出他们的理论后，统计学的基本框架实质上没有发生任何改变。

根据应用最广泛的一种计算方法，如果假设为该现象存在，那么当 *P* 值为 0.01 时，该现象实际并不存在的概率至少为 11%；而当 *P* 值为 0.05 时，这一概率则会上升到 29%。因此，莫德尔的发现是假阳性的概率超过 10%。同样，结果可重复的概率也不是大多数人所想的 99%，而是 73%左右。而再得到一个极为显著的结果的概率只有 50%。换言之，莫德尔的实验结果不可重复的概率高得惊人，我们下面来看一个关于检验故事的案例。

背景：合并 HIV 感染使慢性丙型肝炎患者肝脏疾病的进展加快了 2 倍。HIV 和 HCV 合并感染患者的肝硬化和肝癌的风险显著高于单独 HCV 感染者。另外 HIV 感染降低了丙型肝炎的治疗效果，合并感染者的治愈率比单独 HCV 感染者低 20%。合并感染不仅加快了疾病进程，同时也降低了治疗的应答。无疑合并感染者的死亡率也就更高。主要原因有两方面。一个是 HCV RNA 在合并感染的患者体内更高，而 HCV RNA 的水平是抗病毒治疗应答的影响因素之一。第二个原因是合并 HIV 感染和其他合并症时疾病情况更为复杂，会使用多种治疗药物。不同药物之间会产生相互的影响。由此导致治疗的复杂程度增加，进一步就会降低依从性。注射吸毒致 HCV 与 HIV 合并感染的人群中，此种影响尤为突出。

结论：图 6-1 中横坐标是静脉注射感染率，纵轴是 HIV 合并 HCV 感染率。研究发现静脉毒品注射的 HIV 感染人群与 HCV 感染人群有高度的强关联。

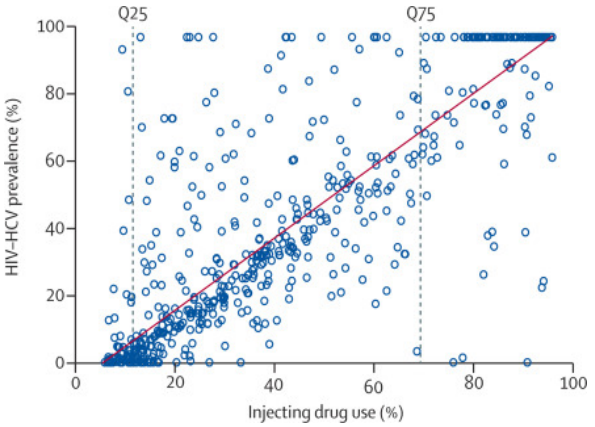


图 6-1　静脉毒品注射感染率与 HIV 合并 HCV 感染率的关联性

资料来源：柳叶刀. LANCET. Prevalence and burden of HCV co-infection in people living with HIV: a global systematic review and meta-analysis.

如图 6-2 所示森林图（Forest plot），可以得到以下主要的内容：

①Odds 代表一个试验结果的可信区间（Confidence Interval，CI），可信区间是真值可能存在的范围，反映结果的准确性。横线越长，说明样本量越小，结果欠准确可靠；横线越短，说明样本量越大，准确性越高，结果越可信。

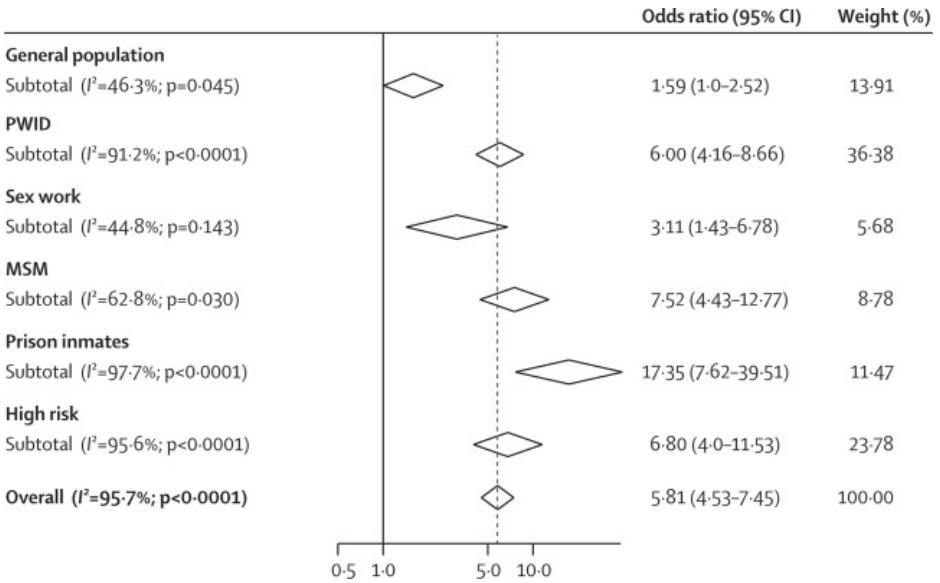
②横线中央的正方形是点估计值，面积大小表示对 Meta 分析的贡献度，即研究权重（Weight）大小。一般来说，对于计数资料如本研究使用的样本量作为权重的衡量依据，样本量越大，权重越大；计量资料则采用标准差作为权重的衡量依据，标准差越小，权重越大。也有以纳入研究的质量评分作为权重的衡量依据。

③最下方的菱形代表多个 RCT 的综合结果。

④垂直线（代表 OR=1，RR=1 或 RD=0）将图分为左右两半，用于判断结果差异有无统计学意义：横线/菱形与垂直线相交则表明该 RCT 中不同治疗措施之间差异无统计学意义。若不相交则有统计学意义，对于不利结局如死亡、疾病进展、残废事件等，横线/菱形完全在垂直线左侧表示治疗组更有效，完全在右侧表示对照组更有效；对于有利结局如治愈、缓解等则相反。

OR 值的全称是 odds ratio，OR 值是相对危险度，又称比值比，对于发病率很低的疾病来说，它的 OR 值即是相对危险度的精确估计值。

如图 6-2 所示，总人群被分为静脉注射、性工作者、男同性恋、监狱囚犯、高风险人群几个组别。荟萃分析比较了 HCV 抗体水平在 HIV 阴性与阳性人群中的不同。比重的百分比采用随机效应分析，其中 *P* 值在不同的人群中也不一样，比如在静脉注射人群中的 *P* 值为 0.0001，这个值统计学意义很显著。就是说假定静脉注射人群中 HCV 伴 HIV 合并感染几率为零，然后用实验数据比对与假设条件吻合的概率，*P* 值越小，推翻假设的概率越大，就是说静脉注射人群合并感染的几率最大。这就是 *P* 值的本质。



资料来源：柳叶刀. LANCET. Prevalence and burden of HCV co-infection in people living with HIV: a global systematic review and meta-analysis.

图 6-2　森林图

6.2 K 线图的故事

K 线图最大的统计意义在于可以同时表达 4 个不同的参数。图 6-3 为一种 K 线图。

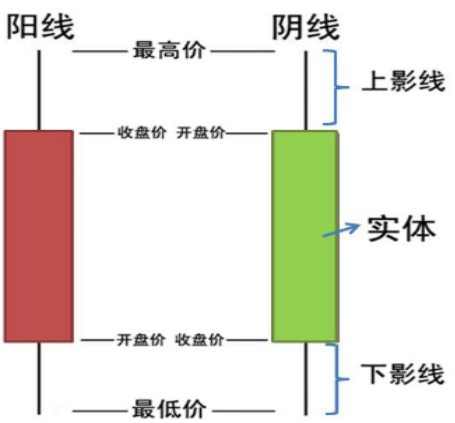


图 6-3　股市上常用的 K 线图

K 线图这种图表源自于日本德川幕府时代，被当时日本米市的商人用来记录米市的行情与价格波动，后因其细腻独到的标画方式而被引入到股市及期货市场。目前，这种图表分析法在我国以至整个东南亚地区均尤为流行。由于用这种方法绘制出来的图表形状颇似一根根蜡烛，加上这些蜡烛有黑白之分，因而也叫阴阳线图表。通过 K 线图，我们能够把每日或某一周期的市况表现完全记录下来，股价经过一段时间的盘档后，在图上即形成一种特殊区域或形态，不同的形态显示出不同意义。我们可以从这些形态的变化中摸索出一些有规律的东西出来。

K 线图形态可分为反转形态、整理形态及缺口和趋向线等。

我们来看看世界级杂志“CELL”中的科技论文对 K 线图的运用方法。如图 6-4 所示

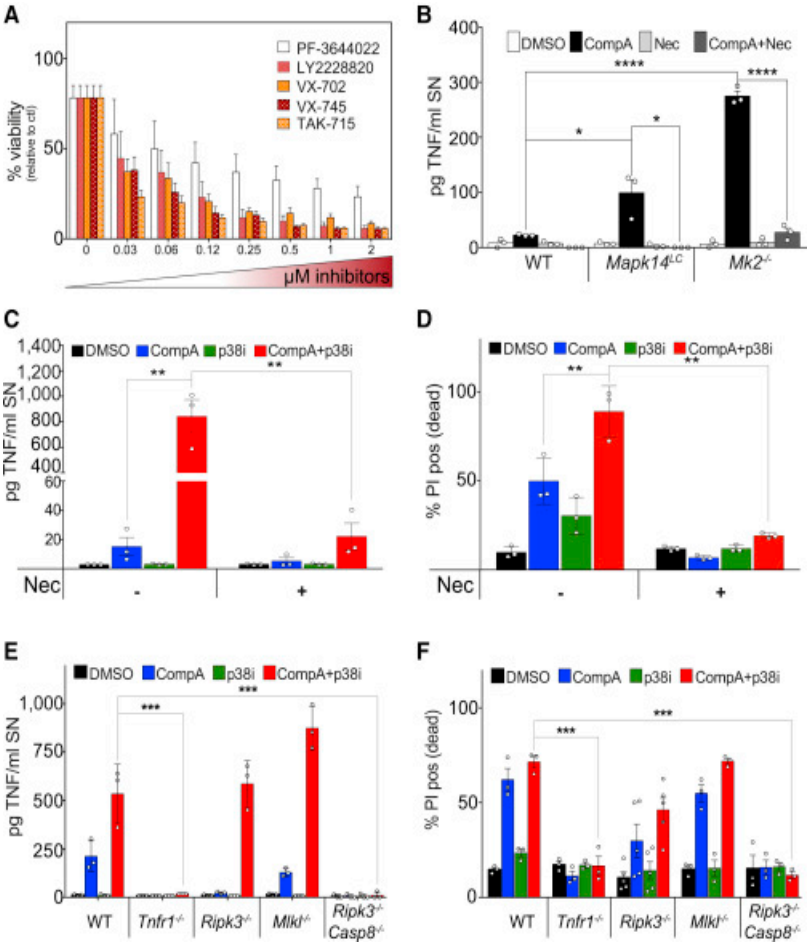


图 6-4　产生抗炎症的 P38 机制

资料来源：CELL. Volume 29,Issue 2,p145–158. 8 February 2016 Targeting p38orMK2 Enhances the Anti-Leukemic Activity of Smac-Mimetics.

（1）基本原理

Smac-mimetics（SM）是一种促使细胞凋亡的靶向药物，经常作为抗癌药物使用在早期的临床试验中。如果能够对其行为模式有一个更深入的了解的话，将会对未来的临床试验设计产生重大的影响。P38/MK2 是已知的产生促肿瘤细胞坏死的炎性细胞因子的主要路径，这一路径可以产生抗炎症的 P38 抑制剂。

意料之外的发现是 P38 抑制剂的使用大大增加了 TNF（肿瘤坏死因子）的生长。在急性骨髓性白血病的治疗中，药物 Birinapant 与 p38 抑制剂的联合使用对抑制癌细胞的生长有很好的疗效。

本研究对 SM 伴 P38 或 MK2 抑制剂的联合用药机理作出了理论模拟与解释。

（2）K 线图的靶向药物机理解读

抑制剂 P38 与 SM 联合用药可以诱发 TNF 的生长，这一生物机制通过激活酶 RIPK1 来实现，而不是通常认为的磷酸化导致细胞程序化坏死的分子机制诱导与激活因子 RIPK3 或 MLKL。

如图 6-4 中 A 图所示，这是一个二维的坐标系，横轴是抑制剂的浓度指标（单位是 mmol/L），纵轴是细胞的活力指标，用百分比表示。实验中的五种抑制剂在不同的浓度条件下对细胞活性的影响程度在 A 图中一目了然，比如 LY2228820 抑制剂随着每单位浓度的增加，细胞的活力百分比也越来越小。A 图表明随着 P38 与 MK2 混合抑制剂的长达 20 分钟的靶向给药，然后再 24 小时 500 个 nM 单位的 companyA 给药后的细胞活性变化情况。

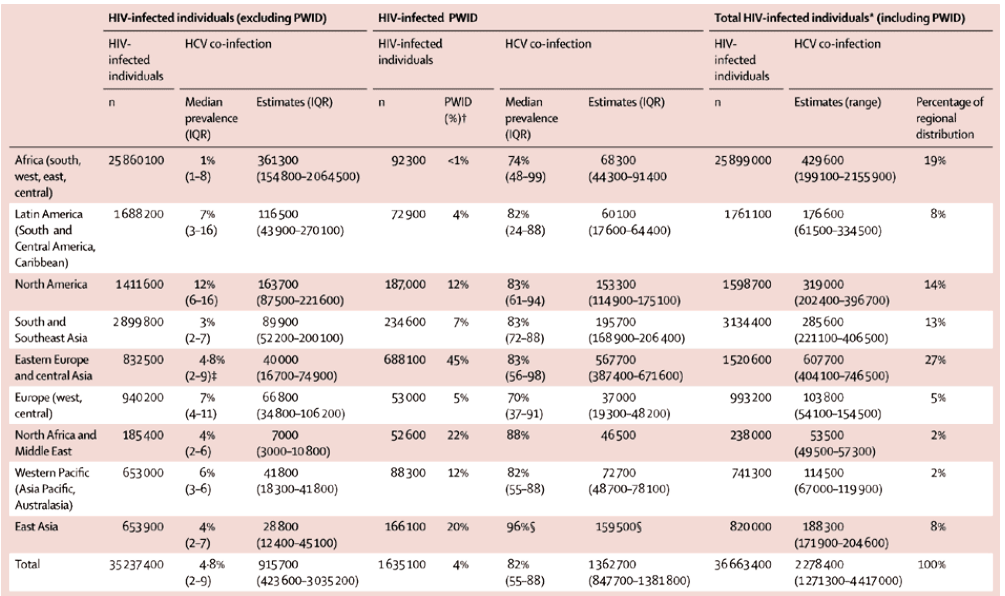
又如图 6-4 中 C 图所示，纵轴是诱发的 TNF 物质量浓度指标，横轴是 NEC 的阴性、阳性指标，NEC 是骨髓非红系有核细胞（nonerythroid cells）的缩写。我们看到四组抑制剂中 CompA+P38i 的组合诱导效果最好，K 线图表明了 TNF（肿瘤坏死因子）的最高、中位、最低的质量浓度指标，从图中可以看出，由于组合抑制剂的使用诱导，CompA+P38i 这一组产生的 TNF 单位质量浓度指标很显著，对阴、阳性的敏感度也显著。

用 K 线图阐述 P38/MK2 作为诱导剂产生 TNF 的分子机制在急性白血病的治疗中有重大的意义，特别是用 K 线图把不同的浓度与质量指标有机地整合就可以用数据的方式完整再现实验的整个过程。特别是肿瘤的靶向治疗与用药的定量分析息息相关，不同的浓度产生不同的结果，阴性与阳性之间，有效与无效之间的差距就是定量实验。定量分析的本质就是浓度、强度、质量、时间长短的数据化与图形化。

K 线图的好处就是在二维的坐标中可以同时表达一个向量的高、中、低不同的组数指标，这样的效果通常要在一个多维坐标中才能图形化，所以 K 线图也可以看成是一个降维的数据表示工具，空间维度与数据的关系再一次昭然若揭。

6.3 国际顶级期刊上的数据技术

如图 6-5 所示，本案例中数据总量为 3523 万条，分为 HIV 感染组（不含静脉注射感染人群）、HIV 感染组（含静脉注射感染人情）和全体 HIV 感染组。采用的基本语如表 6-1 所示。



	HIV-infected individuals (excluding PWID)			HIV-infected PWID				Total HIV-infected individuals* (including PWID)		
	HIV-infected individuals	HCV co-infection		HIV-infected individuals		HCV co-infection		HIV-infected individuals	HCV co-infection	
	n	Median prevalence (IQR)	Estimates (IQR)	n	PWID (%)†	Median prevalence (IQR)	Estimates (IQR)	n	Estimates (range)	Percentage of regional distribution
Africa (south, west, east, central)	25 860 100	1% (1–8)	361 300 (154 800–2 064 500)	92 300	<1%	74% (48–99)	68 300 (44 300–91 400	25 899 000	429 600 (199 100–2 155 900)	19%
Latin America (South and Central America, Caribbean)	1 688 200	7% (3–16)	116 500 (43 900–270 100)	72 900	4%	82% (24–88)	60 100 (17 600–64 400)	1 761 100	176 600 (61 500–334 500)	8%
North America	1 411 600	12% (6–16)	163 700 (87 500–221 600)	187 000	12%	83% (61–94)	153 300 (114 900–175 100)	1 598 700	319 000 (202 400–396 700)	14%
South and Southeast Asia	2 899 800	3% (2–7)	89 900 (52 200–200 100)	234 600	7%	83% (72–88)	195 700 (168 900–206 400)	3 134 400	285 600 (221 100–406 500)	13%
Eastern Europe and central Asia	832 500	4·8% (2–9)‡	40 000 (16 700–74 900)	688 100	45%	83% (56–98)	567 700 (387 400–671 600)	1 520 600	607 700 (404 100–746 500)	27%
Europe (west, central)	940 200	7% (4–11)	66 800 (34 800–106 200)	53 000	5%	70% (37–91)	37 000 (19 300–48 200)	993 200	103 800 (54 100–154 500)	5%
North Africa and Middle East	185 400	4% (2–6)	7000 (3000–10 800)	52 600	22%	88%	46 500	238 000	53 500 (49 500–57 300)	2%
Western Pacific (Asia Pacific, Australasia)	653 000	6% (3–6)	41 800 (18 300–41 800)	88 300	12%	82% (55–88)	72 700 (48 700–78 100)	741 300	114 500 (67 000–119 900)	2%
East Asia	653 900	4% (2–7)	28 800 (12 400–45 100)	166 100	20%	96%§	159 500§	820 000	188 300 (171 900–204 600)	8%
Total	35 237 400	4·8% (2–9)	915 700 (423 600–3 035 200)	1 635 100	4%	82% (55–88)	1 362 700 (847 700–1 381 800)	36 663 400	2 278 400 (1 271 300–4 417 000)	100%

图 6-5　HIV 感染组数据总表

资料来源：柳叶刀. LANCET. Prevalence and burden of HCV co-infection in people living with HIV: a global systematic review and meta-analysis.

表 6-1　基本术语

统计专业术语	主要数据挖掘含义
n	n 主要为样本数，比如总样本数是 3523 万条数据，每个数据就是一个感染者或被抽样调查者。非洲为 2586 万条，拉美为 168 万条，东南亚为 289 万条，东欧为 83 万条等
Median prevalence	中位数感染率，比如北美为 12%，东亚为 4%，拉美为 7%，以东亚为例在 2%～7%之间，中位数为 4%
IQR	interquartile range 四分位数的间距，也叫四分差。内距又称为四分位差，是两个四分位数之差，即内距 IQR=高四分位数—低四分位数
Range	全距（Range），又称极差，是用来表示统计资料中的变异量数（measures of variation），其最大值与最小值之间的差距，即最大值减最小值后所得之数据。极差不能用作比较，单位不同；方差能用作比较，因为都是个比率
Estamates	指预测值，这里的预测值基于两种不同的情况：一是基于 IQR 的数据预测，另一个是基于 Range 的数据预测

图 6-6 详解了本案例的数据处理流程与过程，我们用中文诉述如下：

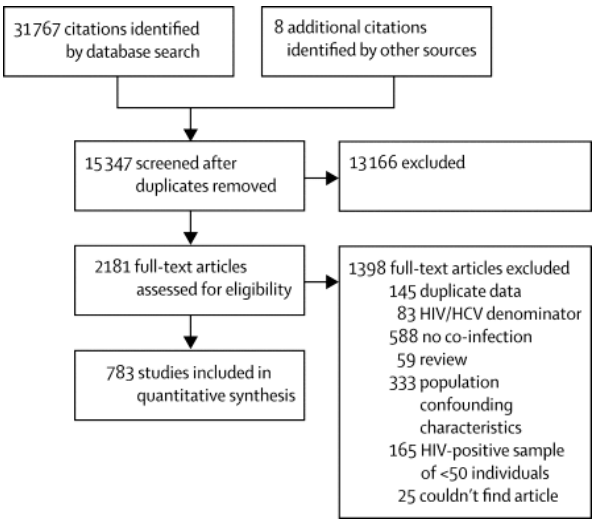


图 6-6　数据处理流程图

① 用数据库搜索的方式确定 31767 篇引文。

② 从其他渠道获取 8 篇引文。

③ 影印文献 15347 篇。

④ 13166 篇被排除。

⑤ 通过合法渠道获取 2181 篇文本文章，其中：1398 篇被排除，145 篇拷贝数据，83 篇合并感染，588 篇没有合并感染，59 篇是老文，333 篇人口信息混乱，165 篇 HIV 阳性者小于 50 例，25 篇文献找不到。

⑥ 783 篇研究中包含综合的定量分析。

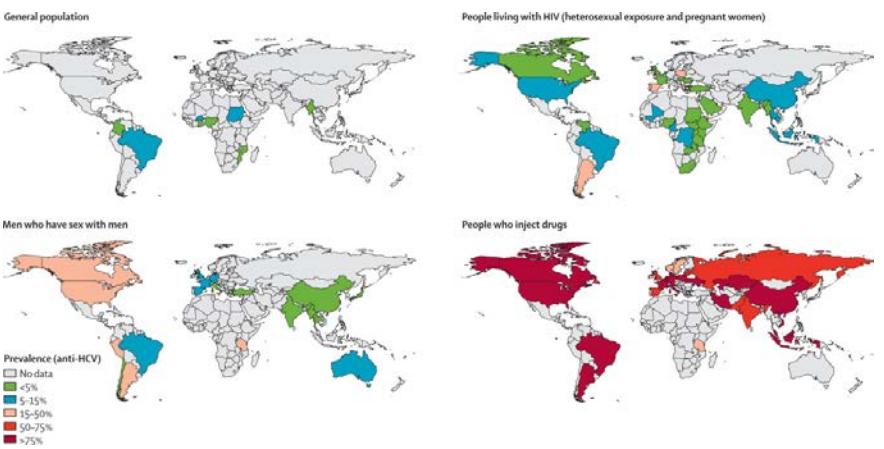


图 6-7　不同人群的地理分布图

背景：小分子核糖核酸基因表达（mirnas）是基因表达至关重要的调节者,对发展、生理过程和真核生物疾病产生广泛的影响。植物和动物 mirnas 的生物起源和功能活动的分子基础高度并行。最近的研究表明,一种潜在的植物 mirnas cross-kingdom 活动,通过饮食摄入,调节哺乳动物基因的表达。尽管在哺乳动物标本中发现植物 microrna 的来源和范围仍有争议，这些初步研究激励我们确定在西方人类血清植物 microrna 能被探测到。这些植物 microrna 是否能够影响基因表达和细胞过程，是否与人类疾病如癌症有关。这里我们发现西方捐赠者血清包含了植物 microrna 的 miR159，其血清丰度与乳腺癌发病率和进展的患者呈负相关。在人类血清，miR159 主要在胞外囊泡中检出,并对过碘酸钠氧化耐受表明了植物来源的末端核糖。在乳腺癌细胞而不是良性乳腺上皮细胞，合成模拟 miR159 能够靶向 TCF7 抑制癌细胞增殖，编码 Wnt 信号转录因子，导致 MYC 蛋白质含量下降。口服 miR159 模拟物显著抑制异种移植乳腺癌小鼠肿瘤的生长。这些结果首次表明，哺乳动物植物 microrna 能抑制肿瘤的生长。

原理：RNAs，包括 miRNAs 的修饰或降解可能在调控 RNA 功能中起着至关重要的作用。多聚腺苷酸化和外来体介导的 RNA 衰减涉及植物 RNAs，包括原始 miRNA 加工中间体的降解。进一步的研究发现，miRNAS 在调控基因表达中有重要的作用，它的存在与 EV（细胞外囊泡）有重要的关系。事实上，miRNAS 寄生在 EV 中常常在人类的血清中可以检测出来。试验证明，mir159 来源于植物中，动物身体中的 RNAS 缺乏修饰性。

结论：植物基因 MiR159 可以在人类血清及乳腺癌组织中被探测到，且与乳腺癌的发展密切相关。

MiR159 在人类血清及乳腺癌组织中都可探测到。如图 6-8 中 A 图所示，miR159 在健康献血者的血清 QPCR 检测（QPCR 的英文全名是 Real-time Quantitative PCR Detecting System，即实时荧光定量核酸扩增检测系统,也叫实时定量基因扩增荧光检测系统,简称 QPCR。）中 n=6（样本为 6），乳腺癌患者化疗后有发展的为 n=10，化疗后没有发展的为 n=20。A 图中横轴有三个组别，分别是发展组、未发展组和健康组。我们可以看到随着血清中 miR159 水平增加，乳腺癌的进展也受到了很大的抑制，特别是化疗后乳腺肿瘤没有发展的那一组中 miR159 水平已经恢复到了健康人群的水平。这从统计学与数据挖掘的角度说明了 miR159 与乳腺癌的关系密切。

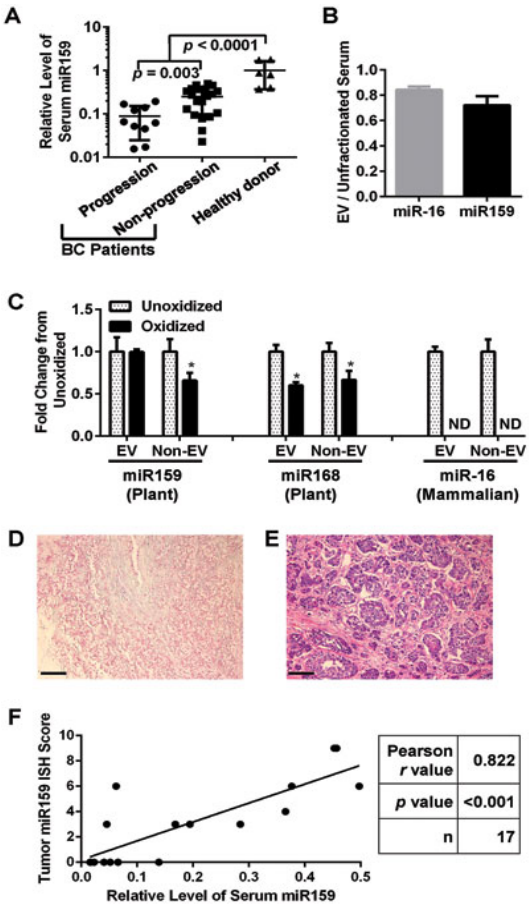


图 6-8　乳腺癌中发现植物基因 MiR19

资料来源：CELL Cross-kingdom inhibition of breast cancer growth by plant miR159

图 6-8 中的 F 图更有统计学分析意义。F 图中横轴为血清 miR159 的相对水平，纵轴为肿瘤 miR159 的 ISH 评分。在统计学中，皮尔逊相关系数（Pearson correlation coefficient）通常用 r 或是 ρ 表示，是用来度量两个变量 X 和 Y 之间的相互关系（线性相关）的，取值范围在[−1,+1]之间。F 图中显示皮尔逊 R 值为 0.822。

更为有意义的是图 6-8 中 A、B、C、D、E、F 图有效地模拟并还原了试验的整个过程，这就是数据技术与生物医学结合的产出。依靠数据图表，各项指标栩栩如生，这就是数据挖掘的强大力量，不仅仅如此，新的知识发现开启了基础医学研究的新篇章。

本文的 KDD（数据库知识发现）表明，植物基因 miR159 在人类血清及乳腺肿瘤中的发现开启了乳腺癌治疗的新篇章，人类在乳腺癌的治疗中找到了跨物种的肿瘤抑制剂。

图 6-9 中的 A-M 图通过数据的统计分析较好地体现了对试验过程的再现、对原理的诠释、对试

验的客观真实记录与还原。

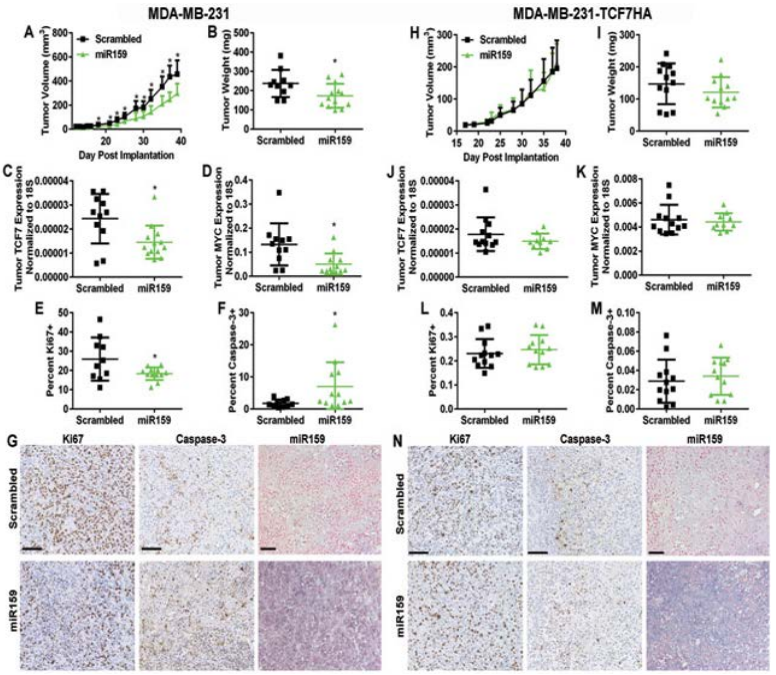


图 6-9　跨物种种瘤抑制剂的机制

资料来源：CELL Cross-kingdom inhibition of breast cancer growth by plant miR159

图 6-9A、B 图中以 MDA-MB-231 乳腺癌细胞株为例，展现了使用 MiR159 组别与对照组在肿瘤细胞株重量与体积方面的差异。A 图中 mir159 植入 40 天后对比最为显著，mir159 组体积只有 200mm^3，而对照组已经有 400mm^3。B 图中即使以中位数为例也可以看到 mir159 组别的肿瘤株重 180mg，比较组已经有 240mg.毫无疑问，就肿瘤株的体积与重量而言，mir159 对肿瘤生长的抑制效果已经用统计数据加图表的办法表现得客观而生动。

既往的科学发现 AF1q 基因是 TCF7/Wnt 信号通路中的关键蛋白，其可以控制癌细胞的行为，该基因表达的增加会促进肿瘤细胞的发育和生长。利用 TCF7 控制乳腺癌基因的表达就是本案例的一个主要试验方向。

myc 基因是较早发现的一组癌基因,包括 C – myc、N –myc、L - myc ，分别定位于 8 号染色体、2 号染色体和 1 号染色体。结构上由不编码蛋白质的第 1 外显子和编码蛋白质的第 2、3 外显子构成。除了染色体易位可破坏 myc 基因的表达调控之外，在某些肿瘤类型中 myc 基因还受 DNA 扩增的影响。myc 基因在小细胞肺癌中有较高频率扩增，在很多其他类型上皮癌如乳腺癌和结直肠癌中也有扩增。

图 6-9 的 C、D、J、K 图分别对控制癌症基因的 TCF7 表达与 MYC 表达作出了比对。以中位数

为例，mir159 组别在控制方面较比较组都有较大的比较优势。

目前，在一张二维坐标系中涵盖多维度的数据信息是国际期刊的重要发展方向。以图 6-9 中的 D 图为例，其数据维度如表 6-2 所示。

表 6-2　D 图中的数据维度

	维度名称
1	肿瘤的 MYC 表达
2	恢复至 185
3	肿瘤基因表达的指标（纵轴）
4	对照组高位数
5	对照组中位数
6	对照组低位数
7	本组高位数
8	本组中位数
9	本组低位数
10	本组、对照组

6.4 SCI 荟萃分析中的统计学工具

本节中纳入研究的文献信息如表 6-3 所示。

表 6-3　纳入研究的文献质量评价

文　献	研究设计			可比性			结果评估			得分
	1	2	3	4	5	6	7	8	9	
蒋一鸣等[3]	-	*	*	*	*	*	-	-	*	6
秦千子等[4]	-	*	*	*	*	*	*	*	-	7
汪建平等[5]	-	*	*	*	*	*	*	*	*	8
侯沙尔等[6]	-	*	*	*	*	*	*	-	*	7
蔡光荣等[7]	-	*	*	*	*	*	-	-	*	6
顾晋等[8]	-	*	*	*	*	*	-	-	*	6
李长华等[9]	-	*	*	*	*	*	*	-	*	7
李义等[10]	-	*	*	*	*	*	*	-	-	6
邓大伟等[11]	-	*	*	*	*	*	*	*	*	8
罗森飙等[12]	-	*	-	*	*	*	*	-	*	6
陈云华等[13]	*	*	-	*	*	*	*	-	-	6
杨宝仁等[14]	-	*	*	*	*	*	-	-	-	5

续表

文　献	研究设计			可比性			结果评估			得分
	1	2	3	4	5	6	7	8	9	
Nakagoe 等[15]	-	*	*	*	*	*	-	-	*	6
Chuwa 等[16]	-	*	*	*	*	*	*	*	*	8
Kim 等[17]	-	*	*	*	*	*	*	-	*	7
Marr 等[18]	-	*	*	*	*	*	-	-	*	6

注：1：随机对照试验；2：具有纳入标准；3：样本量≥50；4：年龄；5：性别；6：Dukes 分期；7：并发症发生率；8：病死率；9：5 年生存率。每符合上述一项指标记一个“*”

6.4.1　研究对象及入选标准

检索策略：通过计算机文献检索，以 Medline、Embase、The Cochrane Library 数据库作为己发表国外文献的主要来源，以维普、万方及中国知网数据库作为己发表国内文献的主要来源。并通过文献追溯和人工检索的方法，收集国内外 2001 年～2012 年间公开发表的关于腹会阴联合切除术与低位前切保肛术治疗低位直肠癌的研究文献。检索时间：2012 年 11 月 11 日。外文检索策略：low rectal cancer AND（ sphinctersparing op-eration OR abdominoperineal resection OR low anterior re-section ）AND（ comparative study OR clinical trial OR e-valuation study OR multicenter study ）。中文检索策略：低位直肠癌 AND（ 腹会阴联合切除术 OR 低位前切术 OR 直肠癌保肛术 ）。

文献纳入标准及排除标准。纳入标准：① 2001 年 1 月至 2012 年 11 月公开发表的腹会阴联合切除术和低位前切保肛术治疗低位直肠癌的临床对照研究的中文和外文文献。② 术前均未接受化学治疗和(或)放射治疗，为原发性直肠癌患者。③ 肿瘤距肛缘的距离小于等于 10 cm。④ 对照研究有足够的样本量，单个研究。样本量大于 20。⑤ 腹会阴联合切除术组（ APR 组 ）和低位前切保肛术组（ LAR 组 ）除会阴部切除不同外，腹盆腔手术均行全直肠系膜切除术，其余处理基本相同。

排除标准：① 复发直肠癌、非原发性直肠癌或直肠良性疾病患者。② 肿瘤距肛缘>10 cm 的患者。③ 术前接受放、化疗的患者。④ 文章重复发表。⑤ 文献提供数据不足且与作者联系未能获得原始数据的。⑥ 单个研究样本量<20。

6.4.2　统计学处理

采用 Review Manager 5.2 统计软件对入选的文献数据进行分析处理。根据齐性检验结果选择计算模型，若研究效应量同质，则采用固定效应模型加权合并；若为异质，则采用随机效应模型加权合并。二分类资料计算优势比（ OR 值 ）或相对危险度（ RR 值 ），连续型资料计算加权均数差（ WMD ）、标准化均数差（ SMD ），同时计算出数据的 95%可信区间（ CI ）及 P 值，P<0.05 为差异有统计学意义。

如图 6-10 所示，从漏斗图中看有无数据偏倚关键是看各试验是不是呈以竖线为中心的漏斗状分布，一般样本量大的试验分布在漏斗顶部，集中在竖线附近，而样本量小的试验均匀散在漏斗图的

下部，如有明显不对称，说明可能存在数据性偏倚。而不是简单看竖线两边的试验是不是相等。漏斗图是通过肉眼去看的，存在很大的主观性，尤其在临床试验较少的情况下，难以判断是否存在数据偏倚。图 6-10 中的 5 年生存率分布来看似不存在明显的数据偏倚。

漏斗的作用是排除偏倚之用。小样本所得的离散度较大，因此常处于漏斗图的底部，大样本离散度则较小，因此处于顶部。正常情况下应该是顶小而底大，如果不是这样，则可能存在较大偏倚。

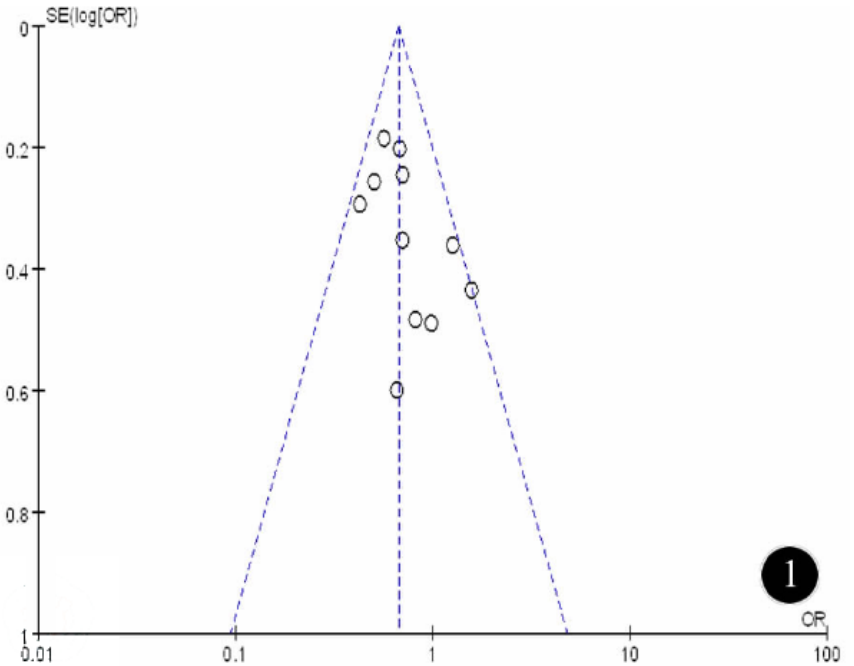


图 6-10　5 年生存率 Meta 分析漏斗图

资料来源：刘德锋 孟翔凌. 腹会阴联合切除术与低位前切保肛术治疗低位直肠癌疗效比较的 Meta 分析. 中华临床医师杂志（电子版）2013 年 5 月第 7 卷第 9 期.

森林图是以统计指标和统计分析方法为基础，用数值运算结果绘制出的图形。它在平面直角坐标系中，以一条垂直的无效线（横坐标刻度为 1 或 0）为中心，用平行于横轴的多条线段描述了每个被纳入研究的效应量和可信区间（Confidence Interval，CI），用一个棱形（或其他图形）描述了多个研究合并的效应量及可信区间。它非常简单和直观地描述了 Meta 分析的统计结果，是 Meta 分析中最常用的结果表达形式。

二分类变量固定效应模型 Peto 法与 M-H 法的优缺点的比较：Peto 法已被广泛应用到临床随机对照试验的 Meta 分析中，利用实际观察值与理论观察值的差别估计效应量的大小，计算方法简单易懂；但 M-H 方法统计效能更强一些，有时 Peto 法会造成效应量的有偏估计，特别是对于队列研究及病例对照研究，两类方法估计结果差别较大，此时将首选 M-H 法；而对于临床对照试验，特别是大规模

的临床随机对照研究，可选用 Peto 法。

图 6-11 如果做亚组分析，与其看亚组是否有意义，不如看亚组之间是否有差异，看 Revma 界面最后一行，Test for subgroup differences 的 *P* 值是否有意义，如果 *P* 小于 0.05，说明两个亚组之间的治疗效果差异很大，也可以说明问题，这就不需要纠结于用何种统计方法。如果行敏感性分析，理由是纳入患者的标准发生了变化。换用不同的统计模型也是敏感性分析的一种。

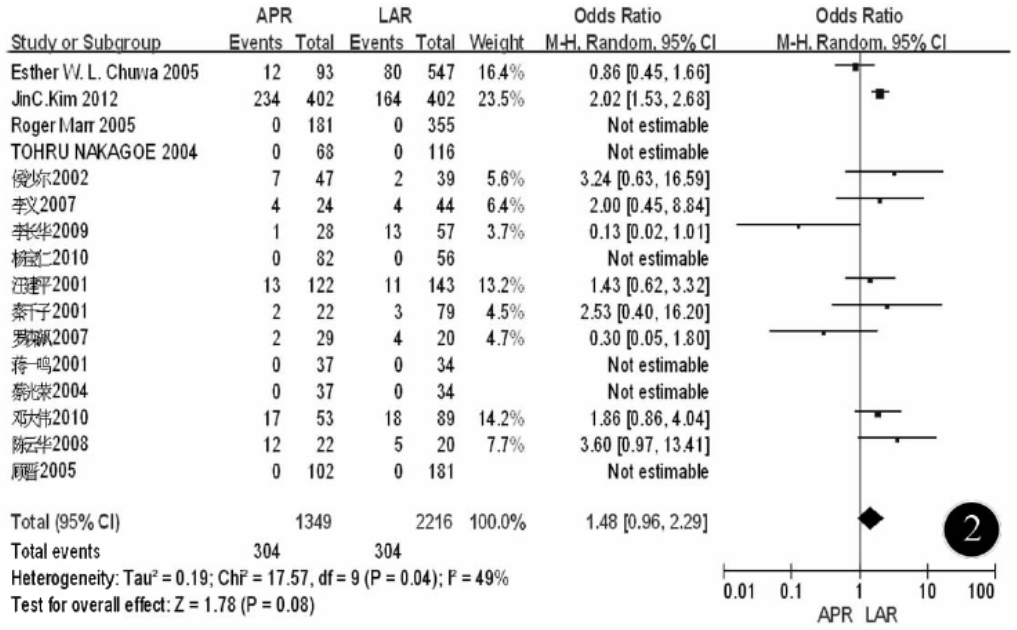


图 6-11　APR 组和 LAR 组的并发症对比的 Meta 分析森林图

RR 和 OR 作为基于随机对照试验中二分类数据的 Meta 分析常用的效应指标，目前在国内外应用中仍存在一些问题，在现有的 Meta 分析报告规范中，尚未明确 RR 或 OR 的选择标准，且无对作者选择原因的说明及报告的要求，作者在 Meta 分析中的指标选择仍具有较大的随意性和难解释性，这可能导致结论的偏倚增大。在二分类随机对照试验的 Meta 分析中，效应指标的不同可能导致结局定量或定性的差异，进行结局事件发生率的预计算和统计指标的修正对指标的选择具有重要的意义。RR 被认为是反映暴露与事件关联强度的最有用指标，而 OR 在结局事件发生率很低时近似于 RR，亦有其应用条件。

结论：原文的结论认为低位前切保肛术治疗低位直肠癌能够获得更好的远期疗效。实际上，医学数据挖掘有完全不同的看法，低位保肛的大量随访数据表明癌症的根治术达到了效果但患者的生活质量严重下降，数据挖掘再一次证明——低位保肛术要慎之又慎。